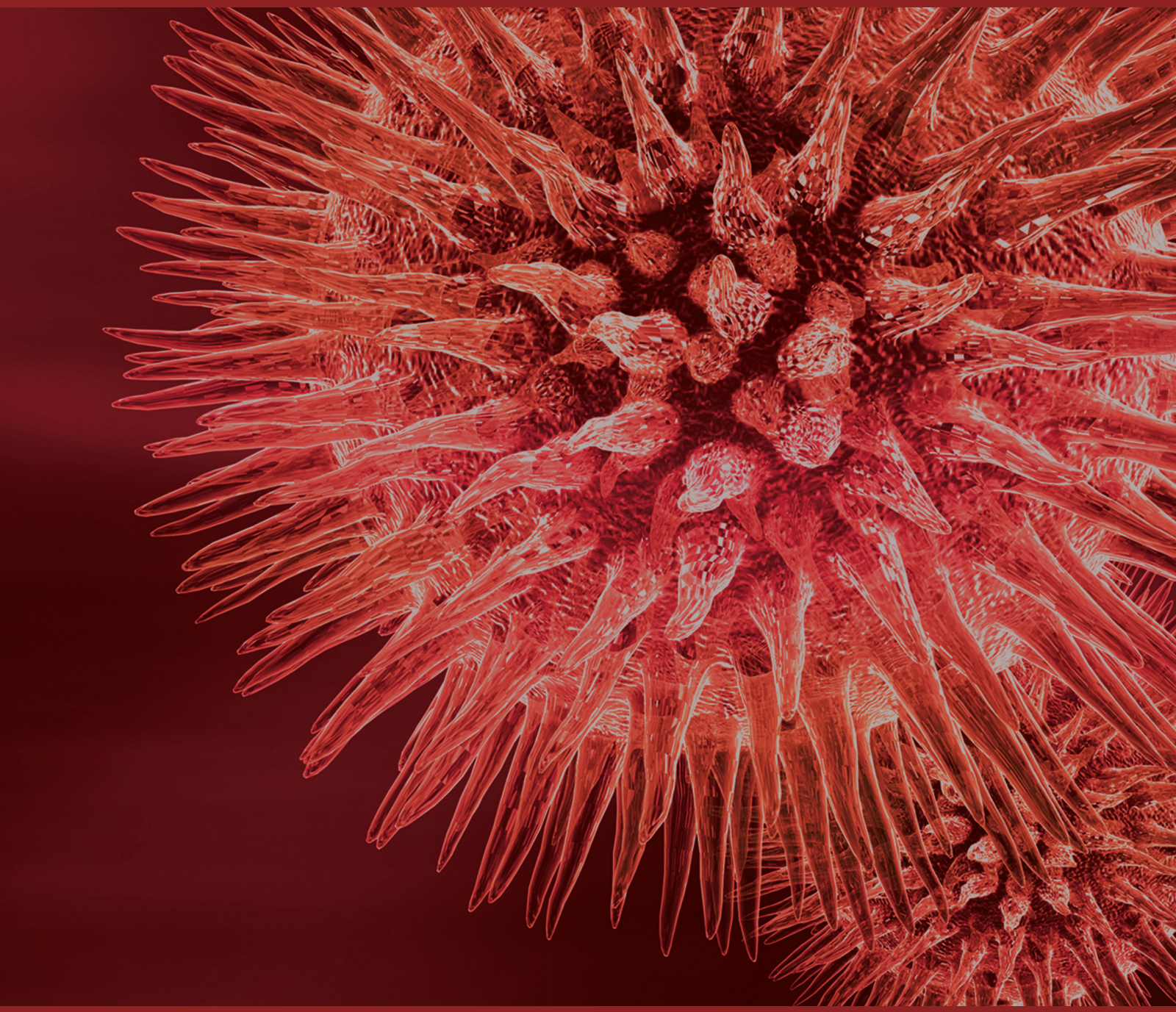
BioMed Research International

Computational and Bioinformatics Techniques for Immunology

Guest Editors: Francesco Pappalardo, Vladimir Brusic, Filippo Castiglione, and Christian Schönbach



Computational and Bioinformatics Techniques for Immunology

BioMed Research International

Computational and Bioinformatics Techniques for Immunology

Guest Editors: Francesco Pappalardo, Vladimir Brusic, Filippo Castiglione, and Christian Schönbach

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 263189, 2 pages
http://dx.doi.org/10.1155/2014/263189

Editorial

Computational and Bioinformatics Techniques for Immunology

Francesco Pappalardo,[1] Vladimir Brusic,[2] Filippo Castiglione,[3] and Christian Schönbach[2]

[1]*Department of Drug Sciences, University of Catania, 95125 Catania, Italy*
[2]*School of Science and Technology, Nazarbayev University, Astana 010000, Kazakhstan*
[3]*Istituto Applicazioni del Calcolo "M. Picone", Consiglio Nazionale delle Ricerche, Via dei Taurini 19, 00185 Rome, Italy*

Correspondence should be addressed to Francesco Pappalardo; fp@francescopappalardo.net

Received 22 July 2014; Accepted 22 July 2014; Published 31 December 2014

Computational immunology and immunological bioinformatics are well-established and rapidly evolving research fields. Whereas the former aims to develop mathematical and/or computational methods to study the dynamics of cellular and molecular entities during the immune response [1–4], the latter targets proposing methods to analyze large genomic and proteomic immunological-related datasets and derive (i.e., predict) new knowledge mainly by statistical inference and machine learning algorithms.

Since immunology provides key information about basic mechanisms in a number of related diseases, it represents the most critical target for medical intervention. Therefore an advance in either computational or bioinformatics immunology research field has the potential to pave the way for improvement of human health through better patient-specific diagnostics and optimized immune treatment.

In this special issue, we take an interest from mathematicians, bioinformaticians, computational scientists, and engineers together with experimental immunologists, to present and discuss latest developments in different subareas ranging from modeling and simulation to machine learning predictions and their application to basic and clinical immunology.

Of the possible directions for development in immune-informatics special interest is raising for models focusing on innate-adaptive immune response activation, immune senescence, and multiscale and multiorgan models of immune-related diseases and for models accounting for cell trafficking in lymph nodes and/or in the lymphatic mesh as in "*Modeling biology spanning different scales: an open challenge*" by F. Castiglione et al.

Exploring the connections between classical mathematical modeling (at different scales) and bioinformatics predictions of omic scope along with specific aspects of the immune system in combination with concepts and methods like computer simulations, mathematics and statistics for the discovery, design, and optimization of drugs, vaccines, and other immunotherapies represents a hot topic in computational biology and systems medicine [5, 6].

The review from F. Castiglione et al. calls attention to the importance of the different time-space scale involved in biological phenomena and in particular in the immune system. It dissects the problem and discusses various techniques that have been developed in scientific areas other than computational biology.

In their paper S. Jarrah et al. illustrate a simple ODE model to investigate the role of the immune response in muscle degeneration and regeneration in the mdx mouse model of Duchenne muscular dystrophy. Their model suggests that the immune response contributes substantially to the muscle degeneration and regeneration processes and predicts in a certain parameter range a permanent immune activation damaging muscle fibers.

In the paper contributed by T. Clancy and E. Hovig, the authors propose a new method to integrate expression profiles and protein-protein interaction (PPI) data. Bioinformatics techniques are used to study differential protein interaction mechanisms across the entire immune cell lineages and the transcriptional activators and modules and are

analyzed in the context of exemplars obtained by clustering the PPI network. The results illustrate that the integration of protein interaction networks with the most comprehensive database of gene expression profiles of the immune cells can be used to generate hypotheses into the underlying mechanisms governing the differentiation and the differential functional activity across the immune cell lineage.

The development of mathematical models of the immune response allows a better understanding of the multifaceted mechanisms of the defense system. In this scenario, as already introduced in the review from F. Castiglione et al., multiscale approaches play a fundamental role. B. de M. Quintela et al. propose a scheme for coupling distinct models of different scales and aspects of the immune system describing a new model that deals with the inflammation processes. These processes are simulated coupling and ordinary differential equations that are used as a model for the systemic response. The dynamics of various immune cells is shown in the presence of an antigen.

There is a controversy about the relationship between HLA-A2 and Alzheimer's disease. HLA supposedly plays a modifier effect on the risk that depends on genetic loadings. Garcia and Murillo present an in silico method to evaluate this relationship and to reveal genes associated with both the HLA-A2 and Alzheimer's disease. They used experimental knowledge of protein-protein interactions to evaluate the top ranked genes shared by both concepts, previously found through text mining.

With the vast amount of immunological data available, immunology research is entering the big data era. These data vary in granularity, quality, and complexity and are stored in various formats, including publications, technical reports, and databases. In the paper contributed by G. L. Zhang et al., it is clearly stated that the present challenge is to make the transition from data to actionable knowledge and wisdom and bridge the gap between knowledge and application. In their work, the authors present a knowledge-based approach based on a framework called KB-builder that facilitates data mining by enabling fast development and deployment of web-accessible immunological data knowledge warehouses. This technique speeds up rational vaccine design by providing accurate and well-annotated data coupled with tailored computational analysis tools and workflows.

Hepatitis C virus and HIV are rapidly mutating viruses. They have adopted evolutionary strategies that allow escape from the host immune response via genomic mutations. Recent advances in high-throughput sequencing are reshaping the field of immune-virology of viral infections, as these allow fast and cheap generation of genomic data. P. Leung et al. propose a pipeline that allows visualization and statistical analysis of viral mutations that are associated with immune escape. Using next generation sequencing data from longitudinal analysis of HCV viral genomes during a single HCV infection, along with antigen specific T-cell responses detected from the same subject, the authors prove the applicability of these tools in the context of primary HCV infection. The proposed pipeline is a freely accessible collection of tools (see the paper for details).

M. Kenn et al. point the attention on the dynamic variations in the distances between pairs of atoms that are used for clustering subdomains of biomolecules. They draw on a well-known target function for clustering and first show mathematically that the assignment of atoms to clusters has to be crisp, not fuzzy, as hitherto assumed, proving that this method reduces the computational load of clustering drastically, demonstrating results for several biomolecules relevant in immunoinformatics.

In the paper by R. Ribarics et al., molecular dynamics is presented as a valuable tool for the investigation of functional elements in biomolecules. They used several spline models to approximate the overall shape of MHC α-helices. The authors applied this technique to a series of MD simulations of alloreactive MHC molecules that allowed them to capture the dynamics of MHC α-helices' steric configurations. In the paper, they discuss the variability of spline models underlying the geometric analysis with varying polynomial degrees of the splines.

HIV represents a widespread viral infection without cure. Drug treatment has transformed HIV disease into a treatable long-term infection. However, the appearance of mutations within the viral genome reduces the susceptibility of HIV to drugs. In the paper contributed by M. Haering et al., the authors discuss predictions derived from a mathematical model of HIV dynamics. Their results indicate that early therapy initiation (within 2 years after infection) is critical to delay AIDS progression.

Francesco Pappalardo
Vladimir Brusic
Filippo Castiglione
Christian Schönbach

References

[1] S. Motta and F. Pappalardo, "Mathematical modeling of biological systems," *Briefings in Bioinformatics*, vol. 14, no. 4, pp. 411–422, 2013.

[2] D. Alemani, F. Pappalardo, M. Pennisi, S. Motta, and V. Brusic, "Combining cellular automata and lattice Boltzmann method to model multiscale avascular tumor growth coupled with nutrient diffusion and immune competition," *Journal of Immunological Methods*, vol. 376, no. 1-2, pp. 55–68, 2012.

[3] F. Pappalardo, E. Mastriani, P.-L. Lollini, and S. Motta, "Genetic algorithm against cancer," in *Fuzzy Logic and Applications: 6th International Workshop, WILF 2005, Crema, Italy, September 15–17, 2005*, vol. 3849 of *Lecture Notes in Computer Science*, pp. 223–228, Springer, Berlin, Germany, 2006.

[4] F. Pappalardo, M. Pennisi, A. Ricupito, F. Topputo, and M. Bellone, "Induction of T cell memory by a dendritic cell vaccine: a computational model," *Bioinformatics*, vol. 30, no. 13, pp. 1884–1891, 2014.

[5] H. H. Lin, G. L. Zhang, S. Tongchusak, E. L. Reinherz, and V. Brusic, "Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research," *BMC Bioinformatics*, vol. 9, no. 12, article S22, 2008.

[6] E. Karosiene, C. Lundegaard, O. Lund, and M. Nielsen, "NetMHCcons: a consensus method for the major histocompatibility complex class i predictions," *Immunogenetics*, vol. 64, no. 3, pp. 177–186, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 363408, 11 pages
http://dx.doi.org/10.1155/2014/363408

Research Article

Differential Protein Network Analysis of the Immune Cell Lineage

Trevor Clancy[1] and Eivind Hovig[1,2,3]

[1] *Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, 0310 Oslo, Norway*
[2] *Biomedical Research Group, Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo, 0310 Oslo, Norway*
[3] *Institute of Cancer Genetics and Informatics, The Norwegian Radium Hospital, Oslo University Hospital, 0310 Oslo, Norway*

Correspondence should be addressed to Trevor Clancy; trevor.clancy@rr-research.no

Received 18 April 2014; Revised 28 June 2014; Accepted 12 July 2014; Published 21 September 2014

Academic Editor: Filippo Castiglione

Recently, the Immunological Genome Project (ImmGen) completed the first phase of the goal to understand the molecular circuitry underlying the immune cell lineage in mice. That milestone resulted in the creation of the most comprehensive collection of gene expression profiles in the immune cell lineage in any model organism of human disease. There is now a requisite to examine this resource using bioinformatics integration with other molecular information, with the aim of gaining deeper insights into the underlying processes that characterize this immune cell lineage. We present here a bioinformatics approach to study differential protein interaction mechanisms across the entire immune cell lineage, achieved using affinity propagation applied to a protein interaction network similarity matrix. We demonstrate that the integration of protein interaction networks with the most comprehensive database of gene expression profiles of the immune cells can be used to generate hypotheses into the underlying mechanisms governing the differentiation and the differential functional activity across the immune cell lineage. This approach may not only serve as a hypothesis engine to derive understanding of differentiation and mechanisms across the immune cell lineage, but also help identify possible immune lineage specific and common lineage mechanism in the cells protein networks.

1. Introduction

Recently, the Immunological Genome Project (ImmGen: http://www.immgen.org/) consortium completed the first phase of their objective to generate a comprehensive resource of the gene expression repertoire across all murine immune cells [1–5]. This massive genomics effort is the so-called "*act one*" [4] in the characterization of the molecular circuitry across immune cells. The ImmGen effort has the goal to chart the entire immune regulatory mechanisms of the hematopoietic cell lineage. Rigorous standardization procedures in flow cytometry and gene expression microarrays are applied to build this resource, resulting in a comprehensive database of the gene expression profiles in the murine immune system. This present study is motivated by the possible benefit of auxiliary bioinformatics analysis of the ImmGen resource. The aim of such additional bioinformatics examination of this rich resource is to help catalyze the process of understanding the molecular mechanisms of differentiation and functional activity across the entire spectrum of hematopoietic cells.

As is often the case with global or systems profiling of immune cells, the experimental approach to collate the data often concentrates on either protein interaction networks, using proteomics based approaches, or gene regulatory networks, using gene expression microarrays. Usually, these two key types of molecular data sources are not integrated in a combined analysis [6]. Integrated proteomics and transcriptomics analysis may elucidate the underlying differential molecular functions being driven by the regulatory networks across all immune cells. Additionally, in the few immunological studies that do incorporate a global systems approach to dissect the complex network of relationships in immune cells a limited set of hematopoietic cell lineages are studied.

In this meta-analysis, we provide an integrative bioinformatics characterization of the immune cell specific protein networks associated with the gene expression profiles from the ImmGen resource. The generation of such a protein network perspective on the ImmGen regulatory networks may offer immunologists an opportunity to derive hypotheses, which can be tested experimentally and either confirmed or refined through further *in silico* and experimental analyses.

The bioinformatics integration of protein networks with large gene expression datasets was demonstrated, over a decade ago, to be highly useful in the elucidation of signaling functions [7]. Bioinformatics algorithms have continued to successfully demonstrate the ability to identify key functional relationships through challenging task of integrating transcriptomics and interactome datasets [8, 9]. Such network integration of data has since often proven to be effective in generating hypotheses to study the precise mechanisms in signaling networks which correspond to the observed changes in the gene expression profiles. One example of this approach is the integration of protein interaction networks and gene expression data using tissue specific gene expression profiles, which has proven to be insightful in recent years in predicting aspects of tissue specific cell biology [10]. Similarly, in this study we perform a tissue specific protein network analysis based exclusively on the immune cell lineage. To that end, we employ the resources at the Immunological Genome Project, which consists of 816 gene-expression profiles across the mouse immune cell lineage and has recently been systematically analyzed to generate the gene regulatory circuits for each of the cell types in the immune cell lineage. This latter effort, algorithmically named as OntoGeNet, for the first time allowed for the comprehensive identification of their potential regulatory modules [11]. In this study, we use affinity propagation to describe a protein network perspective of these gene regulatory networks driving the gene expression across the immune cell lineage through integrative bioinformatics approach to compute differential protein networks functions. Affinity propagation was applied to a protein interaction network similarity matrix to integrate the gene targets of the immune cell regulatory network with possible interaction network mechanisms. This bioinformatics approach can be applied to generate insights into the underlying mechanisms governing the differentiation and the differential functional activity across the entire murine immune cell lineage.

2. Material and Methods

2.1. Sources of Mouse Protein Networks. Protein interactions were sourced from 10 integrated protein interaction databases, as organized by the iRefIndex [12], and were extracted from the binary physical protein interactions through download from iRefWeb [13, 14]. The protein interactions, their annotations, and their identifiers are integrated in this resource using the iRefIndex method by mapping protein identifications across the databases, enabling systematic backtracking to establish the nonredundant identity of the interaction partners. A strict filtering process for each protein interaction was applied, whereby we selected only physical binary protein interactions from the iRefWeb that satisfied all of the following criteria: (a) experimentally verified; (b) within the same organism; (c) at least one supporting publication in Medline, and (d) physically binding protein interactions. This resulted in a mouse protein interaction network consisting of 20,200 binary protein interactions.

2.2. Integration of Immune Lineage Regulatory Modules with the Murine Protein Interactome. For the immune-cell lineage specific information, we utilized the ImmGen consortium data set (April 2012 release) which consisted of 816 expression profiles from 246 cell types of the mouse immune system [1]. We used each of the 7965 ImmGen genes assigned to a network module from the OntoGeNet algorithm, as a source to generate a protein network similarity matrix (PNSM) from a pairwise analysis of these genes (described below). Each target of all 81 course regulatory network modules calculated from the OntoGeNet algorithm [11] on the ImmGen resource was integrated into a bipartite network analysis as described below. Of the 7965 genes among the ImmGen modules, 2133 had at least one network partner in the mouse interactome, reflecting the current incomplete state of interactome databases [15]. The range of coverage of each individual module in the interactome is illustrated in Supplementary Figure 1 in the Supplementary Material available online at http://dx.doi.org/10.1155/2014/363408, which plots the percentage of genes in each module holding at least one interaction in the mouse interactome (where 2 modules had zero interactors and many have greater than 25% coverage within the module, with a maximum of 48%).

The cell types analyzed encompass all the main hematopoietic lineages, including stem and progenitor cells, granulocytes, monocytes, macrophages, dendritic cells (DCs), natural killer (NK) cells, B cells, and T cells. The T cells include many important subtypes types of $\alpha\beta$ T cells, regulatory T cells (T_{reg} cells), natural killer T cells (NKT cells), and $\gamma\delta$ T cells. A table of the known global regulators that modulate the gene expression programs of these immune cell lineages, as used in this study, is illustrated in Table 1. These regulators inform the selection of target genes selected based on application of OntoGeNet on the known regulators of the immune cell lineages in mice. OntoGeNet is an algorithm recently developed to achieve the reverse engineering of lineage-specific gene regulatory modules from the ImmGen gene-expression profiles. It has an innovative feature in that it integrates the lineage tree when predicting its gene regulatory networks, in addition to the gene expression activity, such that the module's genes are recapitulated in related cell types.

2.3. Defining a Measure of Protein Interactome Similarity Scores to Generate Protein Network Similarity Matrices. A protein network similarity matrix (PNSM) was built and mapped to the ontogeny of the murine immune cells gene expression profile as catalogued at ImmGen. Each of the 7965 target genes from the gene regulatory network modules in OntoGeNet was mapped to its protein product counterparts

Table 1: Known transcriptional activators of the immune cell lineages.

Immune cell type	Known transcriptional activators
B cells	POU2AF1, PAX5, EBF1, SPIB, SFPI1, FOXP1
Dendritic cells	RELB, CIITA, AHR, SPIB, SFPI1
Granulocytes	CEBPB, NFE2, SFPI1, FOXO3, CEBPE, FLI1
Hematopoietic stem cells	HLF, LMO2, MYC, MYCN, GATA2, MEIS1, E2F6
Macrophage	CEBPA, CEBPB, SFPI1
Monocytes	CEBPB, SFPI1
Natural killer cells	EOMES, TBX21, SMAD3, GATA3
Natural killer t cells	GATA3, ZBTB16
abT cells	TCF7, BCL11B, GATA3, IKZF2, RORC, SMAD7, TOX
gdT Cells	GATA3, SOX13, ID3

in the mouse protein interaction network (of which 2133 had at least one interaction). We calculated a protein interaction similarity index for all these gene pairs, which compose the PNSM. The pairwise calculation of the protein interaction network similarity scores across all genes in the murine protein networks was calculated by using the Simpson similarity score. For any given pair of genes, A and B, their shared interaction partners were calculated as $(N(A) \cap N(B))$ in relation to the degree of $N(A)$ and $N(B)$ of the mouse protein interaction network overall. The Simpson index was then calculated as the proportion of shared protein interactions between the gene pair relative to the degree of the least-connected gene in the murine protein interaction network.

$$\frac{N(A) \cap N(B)}{\min(|N(A)|, |N(B)|)}. \quad (1)$$

For each of these similarity scores, a real-valued matrix S was constructed, in which an entry S_{AB} corresponded to a value measuring how similar gene A was to gene B in the mouse interactome.

The Simpson index captures the proportion of shared interaction partners between each gene pair, relative to the degree of the least-connected gene [16]. The choice of using Simpson index as the similarity score to build the PNSM was motivated by its effective comparison of two diverse sets of gene pairs in the network, and by so doing not penalize pairs which have large differences in their node degree in the interactome network [16]. Such large differences often occur in molecular networks due to the scale free property of these networks, as well established in recent years [17]. Other similarity indices can be used to compare networks, each having their strengths and weaknesses depending on the biological application. The Jaccard index, for example, effectively captures the proportion of shared nodes between the gene pairs. However this is in proportion to the total number of nodes in both genes irrespective of their individual node degree. Other similarity indexes are more suitable to capture communities within networks or predict biological function, and all of these indices have been extensively surveyed recently [16] and their optimal use characterized.

2.4. Affinity Propagation Applied to the Protein Interaction Similarity Matrix (PNSM) to Identify Exemplar Protein Network Signatures. In recent years there has been a great deal of development of methods to detect clusters, modules, or communities in molecular networks [18, 19] and also to predict the interrelatedness of these groups [20, 21]. The strengths and weaknesses of these different methods have been profiled in recent bioinformatics perspectives [22]. In this study affinity propagation (AP) was applied to the PNSM to identify the differential protein interaction network mechanisms associated with each of the 10 immune cell types analyzed [23, 24]. Affinity propagation holds an advantage over other clustering procedures applied to similarity matrices in that the method does not require values to be in a specific range. Additionally, although there are similar methods which compete with the performance of AP applied to smaller networks [25], the AP method performs optimally on large similarity matrices [26] such as the PNSM developed in this study whereby all pairwise datapoints are considered as candidate exemplars (clusters). Additionally, the exact number of clusters does not need to be specified. Furthermore, the choice of using the AP algorithm for clustering the similarity matrices has the advantage over classical clustering methods in the fact that AP can determine the appropriate number k of clusters, depending on a vector of the median of similarities as input preference for all genes in the mouse genome. The suitability of AP applied to the PNSM for the functional analysis of the immune cell lineage in this study may also be delineated by the feature of AP not only to identify clusters but also to capture compressed information summarizing the identified clusters [23, 27].

The affinity propagation procedure analyzed measures of protein interaction network similarity between all pairs of proteins built into the PNSM and simultaneously considered all pairwise comparisons as potential exemplars (exemplars, in this setting, are groups of proteins which have similar interactions partners in the mouse interactome). Real-valued matrix so-called "messages" which passed between each pairwise comparison are exchanged between data points until a high-quality set of exemplars (or clusters) eventually emerge as the algorithm iterates. In affinity propagation, there are two different types of messages exchanged between each pairwise similarity in the PNSM: "availabilities" $(a(A, B))$ and "responsibilities" $(r(A, B))$. The "availability," sent from candidate exemplar gene B to data point gene A, is a reflection of how each gene B is suitable to be available for gene A to become an exemplar cluster. The "responsibility," sent from data point gene A to candidate exemplar gene B, is a reflection of how each gene A is suitable to serve as exemplar in B. The PNSM, a real-valued matrix S_{AB}, $s(A, B)$ described above, is taken as input for the affinity propagation algorithm. Each data point is assessed for its suitability to be a candidate exemplar. The details of the updating functions computed as affinity propagation iterates are described extensively by Frey

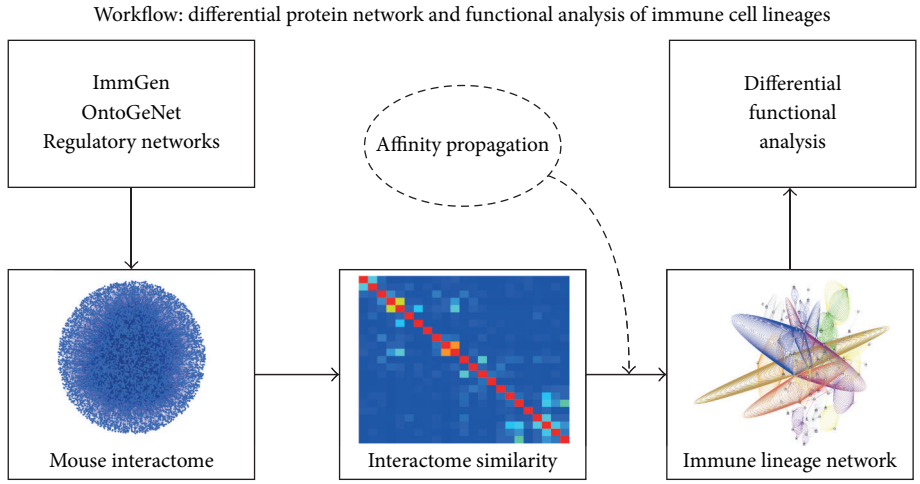


Figure 1: Workflow of affinity propagation on the PNSM and the differential network analysis. The known activator genes which drive the differentiation of the main immune cell types (see Table 1) were used to query the list of their target genes as computed from the OntoGeNet algorithm on the ImmGen resource. The protein network neighborhood of each of the 7965 genes assigned to an ImmGen network module of was integrated with their target lineage information, computed from an integrated set of validated protein interaction network databases. Then, using the Simpson similarity index, their PNSM was computed. Affinity propagation or "message passaging" was then applied on the PNSM, to capture features of the immune lineage network. The resulting exemplars computed from the affinity propagation allows for differential functions to be captured through the lineage tree.

and Dueck [23]. Briefly, the availabilities and responsibility functions are computed as log likelihood ratios, reflecting the evidence accumulated iteratively for how well suited each data point may serve as a candidate exemplar. Initially, the availability is set to zero. The responsibility updates are then computed as

$$r(A,B) \longleftarrow s(A,B) - \max\left\{a\left(A,B'\right) + s\left(A,B'\right)\right\}. \quad (2)$$

The above function allows all candidate exemplars in the PNSM to compete for inclusion of a data point. The availability update then accumulates evidence scores from all data points as to whether each candidate exemplar has likelihood of emerging as an optimal exemplar, using the following update function:

$$a(A,B) \longleftarrow \min\left\{0, r(B,B) + \sum \max\left\{0, r\left(A',B\right)\right\}\right\}. \quad (3)$$

This availability update functions sets the availability of a candidate exemplar to the sum of the positive responsibilities, $r(A,B)$; the candidate exemplar receives from all other data points plus the self-responsibility $r(A,B)$. This self-responsibility, $r(B,B)$, is an evidence score that ranks whether gene B is an exemplar based on the input preferences in the procedure. More extensive details of the affinity propagation algorithm are available in the Frey and Dueck paper [23], which described its development. The functional analysis on the computed exemplars was performed based on structured vocabularies from the Gene Ontology project [28] biological process tree, using a combination of DAVID functional association tools [29] and the gene set enrichment analysis, through the use of GSAT [30, 31]. The bipartite networks, which illustrate the properties of the computed protein networks exemplars with activators of immune cell differentiation, the groups (modules) of coexpressed genes from ImmGen, and the immune cell types, were visualized using the Cytoscape network analysis toolkit [32].

3. Results and Discussion

3.1. Affinity Propagation on Protein Network Similarities, Informed by the Immune Cell Lineage. A protein interaction network similarity matrix (PNSM) was built by computing the similarity in protein interaction partners for all pairwise combinations of genes assigned to modules in ImmGen having been identified as targets of the transcriptional activators driving immune cell lineages as computed from the OntoGeNet algorithm. Affinity propagation was applied on the PNSM in order to identify groups of genes (or "exemplar" clusters), which correspond to similarity in protein interaction network partners. The outcome of this particular approach is the computation of exemplars (or clusters) of genes that share similar protein interaction partners in the mouse interactome. Integrating this level information with the gene regulatory network information captured by OntoGeNet [11] using the ImmGen resource allowed for the capturing of common functional mechanisms among the immune cell lineage. The workflow of the approach applied to achieve this bioinformatics integration of diverse datasets in systems immunology is described in Figure 1. The outcome of the workflow defined by this strategy is illustrated in Figure 2(a). The resulting exemplars (characteristic groups of proteins which have similar interaction partners in the mouse interactome) computed by the affinity propagation are both illustrated in Figure 2 and listed in Supplementary Table 1. It is evident that exemplars with diverse functions were captured using the approach. The nature of this diversity was also

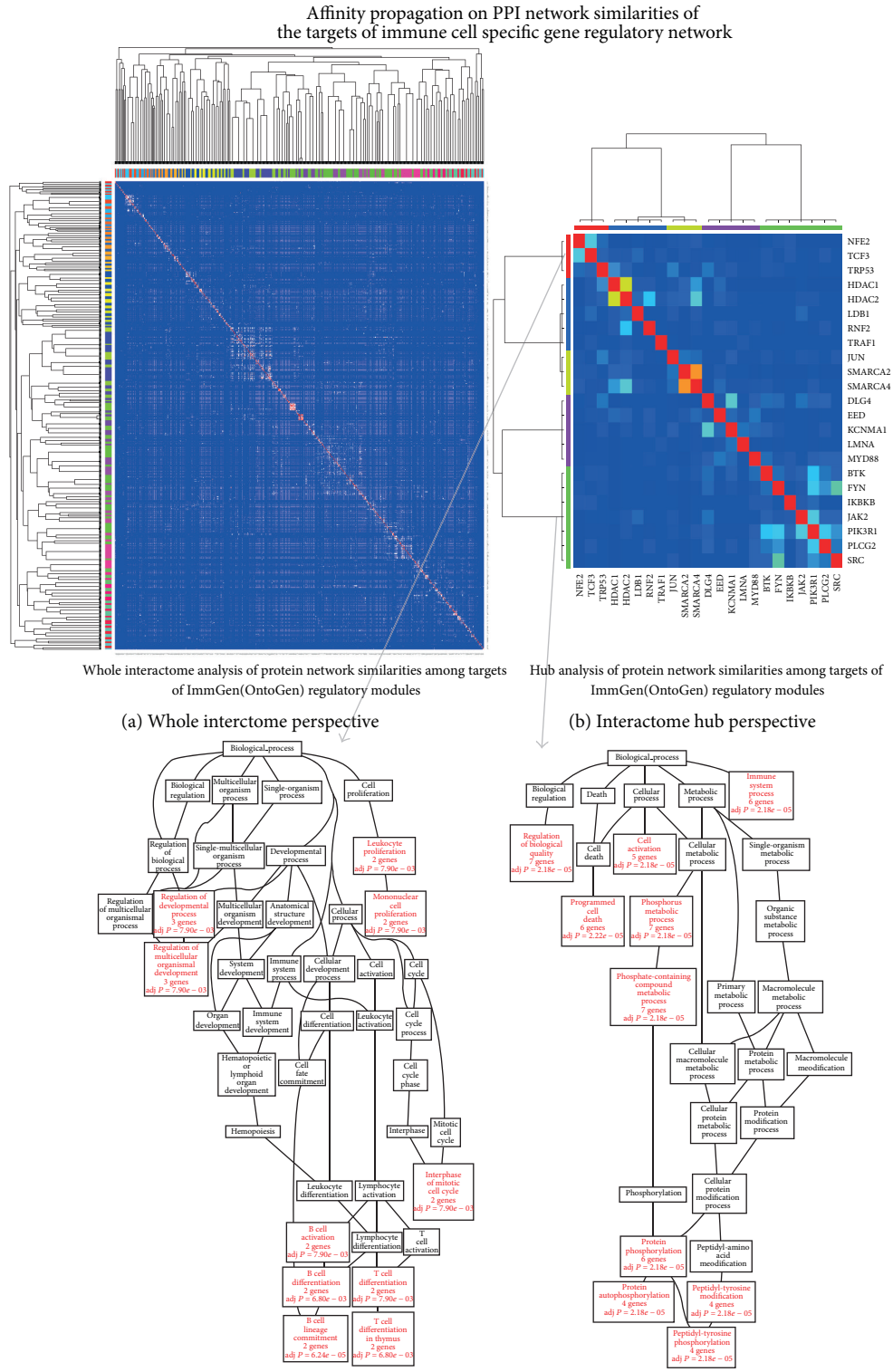


FIGURE 2: Affinity propagation clustering of the protein interaction association similarity matrix of OntoGeNet target module genes. (a) The PNSM is illustrated for both the entire list of target genes computed from the OntoGeNet algorithm on the ImmGen resources. The degree of red color in the heatmap corresponds with the strength of similarity in the protein network for each gene pair. The exemplars as computed from affinity propagation are illustrated in the annotated color bars and the resulting hierarchical clustering (see Table 1 for list of the protein network exemplars). (b) Exemplifies the effect of application of the affinity propagation workflow applied to protein interaction networks, on hubs only. The hub analysis highlights the possible most influential interaction mechanisms activated by the gene regulatory networks (OntoGenet), which govern the immune cell lineage. (c) A differential functional analysis using the Gene Ontology Biological Process (GO-BP) tree is illustrated for two of the computed exemplars. The trajectory of functional significance of GO-PB terms from the two exemplar's genes from Figure 2(b) (indicated by the arrows) is illustrated through the GO-BP tree. GO-BP terms significant for the gene list within the exemplars are highlighted in a red color.

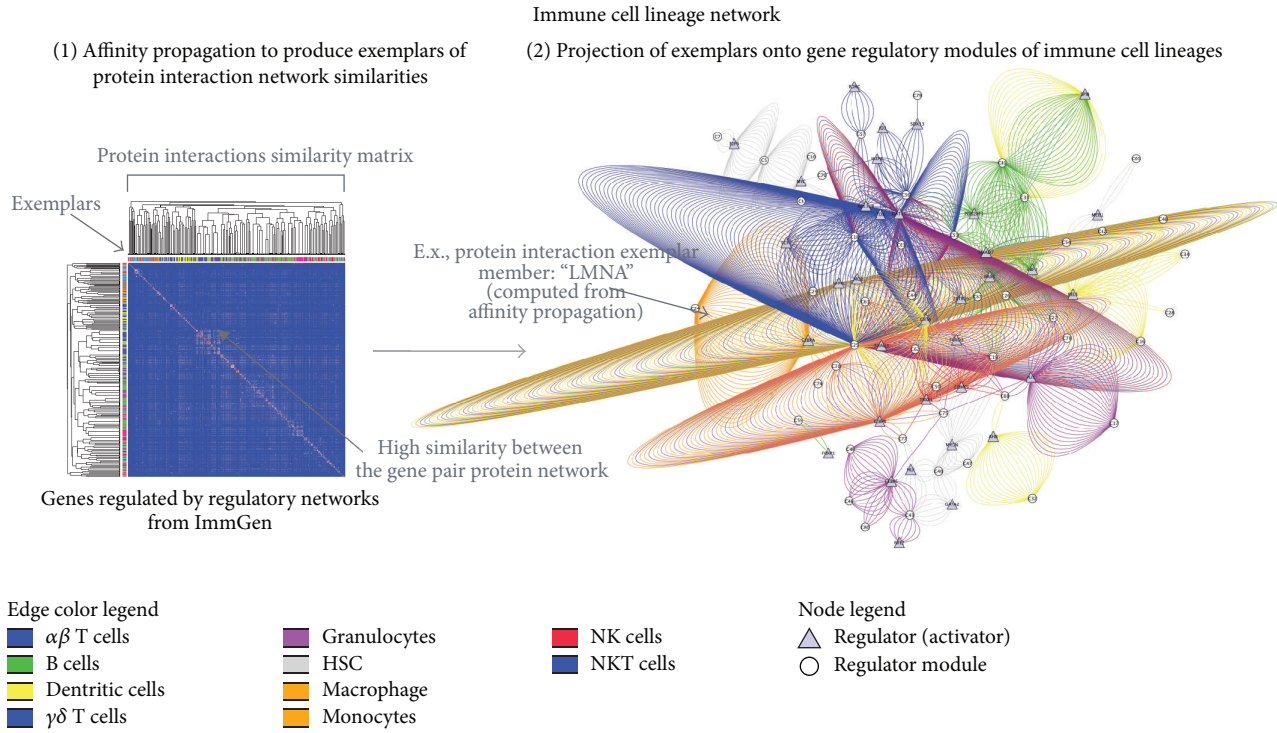


Figure 3: Integrated protein interaction networks perspective on the gene regulation networks driving immune cell lineages. The immune cell lineage network is depicted as a bipartite network with multiple edges representation. Each edge represents a protein network exemplar. The multiple edges connect the different node types and reflect the regulator activity superimposed on the multiple protein network exemplars activated by immune lineage regulators. Each relationship is a representation of the gene regulation modules from the ImmGen resource connecting with the known regulators of immune cell lineages. Each edge in the network represents a relationship between an immune cell line lineage type (see legend in Figure 3(1)) and one of the known activating factors regulating the differentiation of that lineage (see Table 1 for list of the known activators used). An edge is drawn in the network if there is connection between a regulator gene (triangle node) and a course module (groups of commonly expressed genes) calculated from OntoGenet (circle nodes). The number of lines between a regulator and a module is a measure of how many "protein network exemplars," as calculated from the affinity propagation (see main text), are associated to the regulatory module (and therefore a possible measure of the diversity of signaling networks activated in driving the lineage of the immune cell type).

illustrated using the same workflow implemented on target proteins that are also hubs in the mouse protein interaction network (as depicted in Figure 2(b)). A differential functional analysis using the Gene Ontology Biological Process (GO-BP) tree was also applied using gene set enrichment statistical approaches to analyze some of the resulting exemplars [31]. This differential functional analysis is illustrated for two of the computed exemplars in Figure 2(c). The trajectory of functional significance of GO-PB terms from selected two exemplar's genes in Figure 2(c) highlights immune relevant yet diverse, functional mechanisms captured and possibly implicated as important, in immune lineage differentiation. The two GO-BP functional analyses illustrated in Figure 2(c) highlight significant associations for immune relevant terms, and the different exemplars protein networks which were used to generate these associations have differential functional paths through the GO-BP process tree.

3.2. Diverse Quantity and Type of Protein Network Exemplars Associated to Gene Expression Programs among the Immune Cell Lineage. In Figure 3, a bipartite representation of the affinity propagation on the PNSM is depicted. This bipartite network of protein network exemplars and OntoGeNet gene expression modules driven by well-known activators of immune lineage gene expression (Transcription factors from Table 1) is illustrated using multiple edges, whereby each edge is color coded to represent an exemplar as computed using the affinity propagation algorithm on the PNSM. This immune lineage network in Figure 3 illustrates how certain specific clusters in the protein similarity network (exemplars) are potentially activated in specific immune cell types, as indicated by the single color of the respective edges representing the protein network exemplars, while others are potentially activated in multiple lineages during immune cell differentiation, as indicated by many colors of the respective local network. From the strategy employed in this study, as outlined in Figure 1, it is apparent, for example, that there are possibly many distinct protein interaction network mechanisms (as represented by the multiple exemplar edges) associated in the bipartite network with the T cell lineage (where the protein network exemplars are color coded as blue edges in the network), some of which are also shared with natural killer (NK) cells (where the protein network

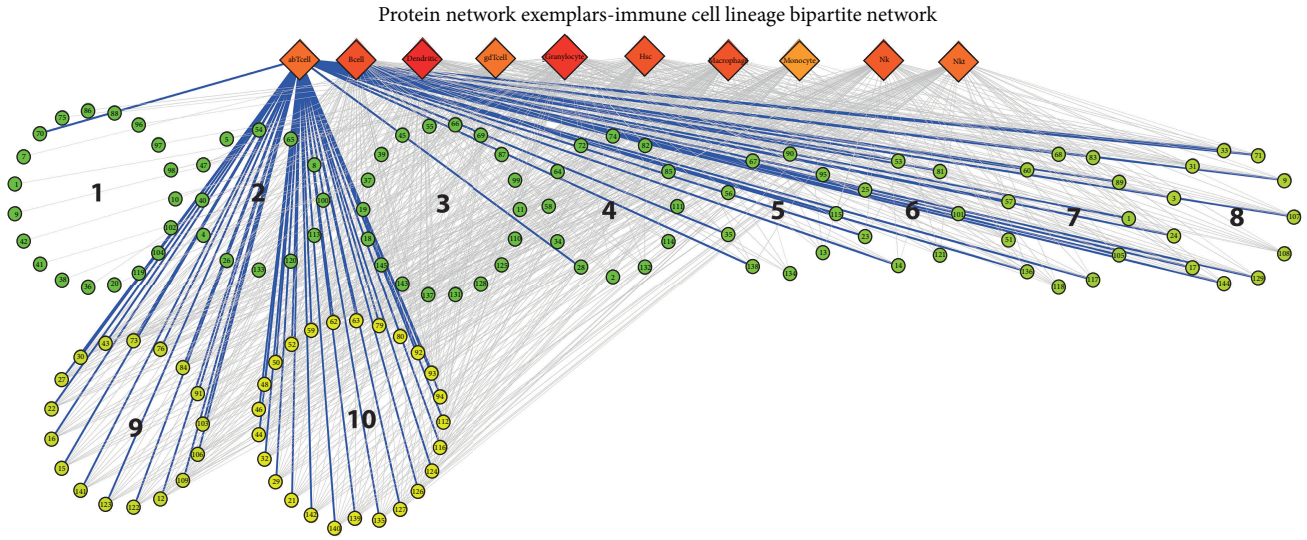


Figure 4: Bipartite network representation of immune cell lineages and protein network exemplars. In this bipartite network representation, the protein network exemplars are represented as circles (the names and genes in these exemplar groups are listed in Table 1), and immune cell types represented as square diamonds. A color gradient of degree of the node in the bipartite network ranges from green (lowest) to yellow (intermediate) and red (highest) is represented. Lineage specific exemplars are clearly illustrated in addition to the increasing range of common protein network exemplars. (These two types of patterns are listed in Table 2 and 3, resp.). The protein network exemplars are ordered, 1–10, according to their connectivity to the ten immune cell types in the bipartite network.

exemplars are color coded as red edges in the network). This NK cell and T cell example are indicative of shared protein network functions in these two different immune cell lineage types, which may be implicated in their shared cytotoxic abilities and immune cell effector functions. The observation is interesting when considering that although NK cells are not generated in the thymus, they share some key molecular characteristics and protein interactions with T cells. For example, they both have some common surface markers. Additionally, NK cells also use the same generic killing mechanisms as cytotoxic $CD8^+$ T cells, although NK cells do not have rearranged T-cell receptor molecules and therefore belong to the innate immune system. Although being regulated by different activators, most of the immune cell types appear to be activated by some common mechanisms with the cells protein network (as illustrated by the activator to module relationships hosting protein network exemplars associated to multiple colors representing all the immune cell types).

Additionally, the immune cell lineage network in Figure 3 also illustrates protein network exemplar relationships that possibly conform to immune cell lineage specific functions (as illustrated by the activator to module relationships hosting protein network exemplars associated to single color representing the immune cell type). This is exemplified in the exemplars specific to the hematopoietic stem cell (HSC) lineage depicted as grey edges in the bipartite network in Figure 3. Here, the protein network exemplars activated in the HSC immune cell types are not activated in others in the lineage network, indicating a possible diminishing importance or deactivation of these functions as immune cells terminally differentiate beyond the HSC lineage.

3.3. Functional Similarity from a Lineage Specific and Common Lineage Perspectives. A different node-type of bipartite network representation of the protein network exemplars with the ten different immune cell types are represented in Figure 4. In this network, the two nodes categories are immune cell types and the calculated protein network exemplars. It is evident that the strategy of applying affinity propagation on the PNSM allows for the capturing of both lineage specific and shared protein network exemplars with diverse functions implicated. These clusters correspond to groups of proteins whose interactions in the cell are possibly more important for the functional activity of lineage specific and common immune cell lineage functions. Such bipartite network representations of protein network exemplars and immune cell types may serve as useful descriptions of both common and more specialized protein interaction functions among immune cell types. In Table 2, an overview of the range of lineage specific protein network exemplars illustrated in the bipartite network in Figure 4 is provided, with some indication of their functional associations. Similarly, lists of those protein network exemplars identified as common to all ten immune cell types analyzed are provided in Table 3.

The gene functions associated with the lineage specific exemplars range from a possible activation of a cell specific Jak-Stat pathway in hematopoietic stem cells (HSC) to the lineage specific gene regulation activity in dendritic cells (see Table 2). It is interesting to note that the Jak-Stat pathway is established to be critically important in regulation of the differentiation mechanisms among stem cells. Additionally, mutations in the Jak-Stat pathway are known to cause destabilization of HSC homeostasis and lead

Table 2: Lineage specific protein network exemplars.

Exemplar ID	Genes assigned to the protein network exemplar	Functional annotation	P value	Immune cell type
102	EPOR, PTPN1, STAT5B	Jak-STAT signaling pathway (KEGG)	$5.20E-02$	Hematopoetic stem cells
20	CCNT1, EIF2B1, MYC			Hematopoetic stem cells
86	EZR, NGFRAP1, NTF3	Neurotrophin signaling pathway	$4.50E-02$	Granylocytes
49	GRB7, TIA1			Hematopoetic stem cells
104	ARRB1 BGN PTS			Macrophages
96	PLCB2, POLA1, VIM			Granylocytes
98	POT1A, TERF1	Telomere maintenance via telomerase (GO-BP)	$5.90E-04$	Hematopoetic stem cells
7	AR, ATRX, CTCF, SMC1A, SMC3	Cell cycle (KEGG)	$4.40E-02$	Dendritic cells
119	CRE, PRPF40A, SMC2	Nucleoplasm (GO-CC)	$8.00E-02$	Dendritic cells
61	CLDN11, CNOT6L, ITGA5, ITGB1, SPARC	Cell adhesion molecules (CAMs) (KEGG)	$7.80E-02$	Macrophage
10	BICC1, CREBBP, CSK, KHDRBS1, PRMT1, RBM39	Control of Gene Expression by Vitamin D Receptor (KEGG)	$7.00E-02$	Ggranylocyte
97	EIF3L, POLR1B, POLR1E	RNA polymerase (KEGG)	$4.70E-03$	Hematopoetic stem cells
41	BRCA2, CEBPD, FANCD2	Cell cycle process (GO-BP)	$5.70E-02$	Macrophages
70	ADRB2, DLL1, MAGI3	Plasma membrane (GO-CC)	$5.40E-02$	abTcells
88	CHORDC1, IGBP1, NR3C1, PPP5C	Transition metal Ion binding (GO-MF)	$3.90E-02$	Hematopoetic stem cells
75	MEOX1, MEOX2, TLE4	Transcription factor activity	$3.40E-03$	Granylocytes
42	FBXW7, NOTCH1, NPM1 STAT4	Notch signaling pathway	$1.20E-02$	Hematopoetic stem cells
36	AEBP2, EED, MORC3, SETX, UHRF1	Nucleoplasm (GO-CC)	$2.30E-03$	Dendritic cells
38	EIF3A, EIF3B, EIF3I, EIF4E	Translational initiation (GO-BP)	$2.00E-08$	Hematopoetic stem cells

to many blood disorders [33]. Notably, many of the lineage specific exemplars in supplementary Table 1 and the bipartite network in Figure 4 are associated to HSCs (42%). This is also illustrated in the network of regulatory and exemplar activity in Figure 3. The increasing degree of promiscuity among the protein network exemplars is evident in the bipartite network by the number of connections of an exemplar to different immune cell type ranges from lineage specific associations to increasingly common lineage associations. One such common lineage protein network exemplar listed in Table 3 is cytokine-cytokine receptor signaling, centered on tumor necrosis factor (TNF) superfamily protein interactions (associated to all ten immune cell types in the bipartite network). As expected, such important protein interactions are preserved across all ten of the protein lineages identified, as TNF-TNF receptor signaling mechanisms are critical for the intercellular communication common to immune cell activity ranging from cell proliferation to apoptosis of immune cell populations.

4. Conclusions

In this study, we have described how the integration of protein interaction networks with the most comprehensive database of gene expression profiles of the immune cell lineage (ImmGen) can be used to generate insights into the underlying mechanisms governing the differentiation and the differential functional activity across all immune cell types. To perform this bioinformatics integration efficiently, and at a large scale, we used affinity propagation applied on a similarity matrix of immune lineage targets gene's interaction partners in the mouse interactome (the PNSM). The approach outline here not only may serve as hypothesis engine to derive understanding of differentiation and mechanisms across the immune cell lineage, but also help identify possible immune lineage specific and common lineage mechanism with the protein networks of the various cell types.

The potential value in applying affinity propagation on the mouse PNSM as a viable strategy to characterize protein network clusters important for immune cell function in human studies is a questionable issue, considering the evolutionary distance between mouse and human [34]. It could be argued, therefore, that bioinformatics strategies such as that described in this study may not be directly applicable to human studies which attempt to capture signatures of immune cell activity using protein networks [35]. However, as mouse is an often used and a powerful model organism for human medicine, it will be exciting to assess the impact of this and similar bioinformatics procedures on the inference

Table 3: Common protein network exemplars.

Exemplar ID	Genes assigned to the protein network exemplar	Functional annotation	P value
127	TANK, TNFRSF11A, TNFRSF4, TNFRSF9, TRAFD1, ZBP1	Cytokine-cytokine receptor interaction (KEGG)	$3.50E-06$
32	ATP6V1A, BAIAP2, CAMK2D, CLU, DLG4, DYNLL1, EPS8, FZD1, FZD4, GRIA3, HCK, INADL, KCNJ10, NSF, OPRD1, PACSIN1, PACSIN2, PGK1, PHB2, PPP3CA, RGS12, SEMA4B, SEMA4C, SLC9A3R1, STXBP1, SYT1	Wnt signaling pathway (KEGG)	$1.60E-03$
44	CD2AP, DAPK1, EFNA2, EPHA4, FGR, FYN, PKD2, RAVER1, RGS1, SH2D1A, SH3BP1, SLAMF1, VAV1, VAV2, ZAP70	T cell receptor signaling pathway (KEGG)	26.7
93	HNRNPA1, LGALS3, NCOA3, PIAS1, RNF19A	Transcription cofactor activity (GO_BP)	$5.30E-02$
29	CLNK, FYB, LYN, PECAM1, SKAP2	Leukocyte activation (GO_BP)	$6.30E-02$
62	CSF2RB, EIF2AK2, GHR, GMCL1, GNB2L1, HES1, IFNGR1, IL6ST, JAK1, JAK2, JAK3, LMO4, NCL, PAG1, SLC40A1	Jak-STAT signaling pathway (KEGG)	$2.10E-09$
139	BRCA1, CIITA, GTF2H2, KDM5D, PDLIM4, POLR1A, POLR2B, TRIP4, TRPS1, ZFP111, ZFP292	Zinc ion binding (GO_BP)	$5.30E-07$
135	RIPK2, TAX1BP1, TIFA, TNFAIP3, TNIP2	Apoptosis (GO_BP)	$6.70E-03$
112	ANXA2, BMYC, GAB2, PRKCE, PRMT5, S100A10	Fc epsilon RI signaling pathway (KEGG)	$1.40E-02$
140	CCNG1, CCNG2, E2F1, GSTA4, NR4A3, SWAP70, TRIM32, TRP53	Cell cycle (GO_BP)	$1.60E-03$
124	HSPA5, HSPA9, HSPD1, PDXK, RAPGEF4, STIP1, YWHAE	Adenyl ribonucleotide binding (GO_MF)	$6.60E-04$
79	ALDOA, CD40, IL1R1, IL1RAP, IL1RL1, IRAK3, IRAK4, IRF4, IRF5, LRRFIP1, MYD88, TIRAP, TLR4, TNFRSF13B, TUBA1A	Cytokine-mediated signaling pathway (GO_BP)	$2.50E-07$
126	BCAR1, BLK, CD22, CD247, CLEC7A, ERBB2, IL15RA, ITGB3, RANBP2, SYK, TUBA4A, WIPF1	Cell surface receptor linked signal transduction (GO_BP)	$4.10E-03$
21	CCL3, CCL4, CCR1, CCR3, CCR5	Chemokine signaling pathway (GO_BP)	$9.80E-07$
63	APOE, CASK, CNN3, CTTN, HSP90B1, IPO11, KCNMA1, LDHA, MYO5A, NDEL1, NUDC, PRDX2, RAB6B, ROCK2, SH3BP4, TPM1, TPT1, TRF, TUBB5	Cellular homeostasis (GO_BP)	$9.20E-04$
80	MYO1C, PCDH15, RICTOR, RRN3, VPS35	Cytoskeleton organization (GO_BP)	$9.30E-02$
92	GFR, PDGFRA, PDGFRB, PTEN, SLC9A3R2	Transmembrane receptor protein tyrosine kinase signaling pathway (GO_BP)	$2.80E-06$
142	CSNK1E, NXN, PLCG2, RAD51, VANGL1, VANGL2	Wnt signaling pathway (KEGG)	$3.90E-03$
48	GRB10, IGF1R, MAP3K5	Insulin-like growth factor receptor signaling pathway (KEGG)	$1.30E-03$
50	HCST, IL2RB, KLRK1, TYROBP	Integral to membrane (GO_BP)	$9.50E-02$
52	ANP32A, DACH1, FOSB, HDAC2, HDAC9, L3MBTL2, MTA1, MTA3, RBBP7, REST, SP3, TCF7L2, WDR5, ZDHHC13, ZFPM1	Regulation of transcription (GO_BP)	$4.00E-09$
116	F2RL2, MAP3K2, SMAD1, SMAD5, TOB1, ZEB2	Cell surface receptor linked signal transduction (GO_BP)	$4.60E-01$
59	CASP1, CASP3, CASP8, CEBPB, GZMB, IL1B	Regulation of apoptosis (GO_BP)	$1.30E-05$
46	E2F2, GATA1, GATA3, GFI1B, LMO2, NFYA	Transcription (GO_BP)	$1.30E-03$

Table 3: Continued.

Exemplar ID	Genes assigned to the protein network exemplar	Functional annotation	P value
94	AXL, CD19, EPHA2, IL4RA, INSR, IRS2, KRAS, NEDD9, NME2, PDCD4, PIK3AP1, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PLCG1, PTK2B, RALGDS, RASSF5, SIRPA, SOCS6, TEK, TLR2	Cell surface receptor linked signal transduction (GO_BP)	$2.10E-03$

of activated protein networks exemplars in disease associated hematopoietic cells in mice models. One such powerful application, for example, is that of gene expression profiles tumor specific $CD4^+$ T cells in a mouse model of multiple myeloma [36], which could possibly capture tumor specific protein interaction network mechanism using the approach described here. Additionally, the gene regulatory network programs we used from OntoGeNet and the expression profiles from ImmGen are conserved between mice and human [37]. With that in mind it is likely that many inferences can be made from mouse model to human immune cell biology using the bioinformatics strategy described here. A natural extension, however, is to apply this strategy to infer protein interaction network mechanisms on similar projects to ImmGen from the compendiums of human immune cells. Namely, the Human Immunology Project Consortium (HIPC) is currently developing standards for data collection, integration, data exchange, and development of a central database of systems immunology data in human samples. The bioinformatics approach described may be fruitful when applied to these and similar upcoming large-scale data sets. Such a protein network integration of the immune system's gene expression compendiums of model organisms may help identify protein interaction mechanisms which are shared among immune types linked to their differentiation, in addition to immune lineage specific immunological mechanisms in protein networks.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publishing of this paper.

References

[1] C. Benoist, L. Lanier, M. Merad, and D. Mathis, "Consortium biology in immunology: the perspective from the immunological genome project," *Nature Reviews Immunology*, vol. 12, no. 10, pp. 734–740, 2012.

[2] T. S. P. Heng, M. W. Painter, K. Elpek et al., "The immunological genome project: networks of gene expression in immune cells," *Nature Immunology*, vol. 9, no. 10, pp. 1091–1094, 2008.

[3] G. Hyatt, R. Melamed, R. Park et al., "Gene expression microarrays: glimpses of the immunological genome," *Nature Immunology*, vol. 7, no. 7, pp. 686–691, 2006.

[4] C. C. Kim and L. L. Lanier, "Beyond the transcriptome: completion of act one of the immunological genome project," *Current Opinion in Immunology*, vol. 25, no. 5, pp. 593–597, 2013.

[5] T. Shay and J. Kang, "Immunological Genome Project and systems immunology," *Trends in Immunology*, vol. 34, no. 12, pp. 602–609, 2013.

[6] J. L. Gardy, D. J. Lynn, F. S. L. Brinkman, and R. E. W. Hancock, "Enabling a systems biology approach to immunology: focus on innate immunity," *Trends in Immunology*, vol. 30, no. 6, pp. 249–262, 2009.

[7] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, supplement 1, pp. S233–S240, 2002.

[8] I. Ulitsky and R. Shamir, "Identification of functional modules using network topology and high-throughput data," *BMC Systems Biology*, vol. 1, article 8, 2007.

[9] I. Ulitsky and R. Shamir, "Identifying functional modules using expression profiles and confidence-scored protein interactions," *Bioinformatics*, vol. 25, no. 9, pp. 1158–1164, 2009.

[10] A. Bossi and B. Lehner, "Tissue specificity and the human protein interaction network," *Molecular Systems Biology*, vol. 5, article 260, 2009.

[11] V. Jojic, T. Shay, K. Sylvia et al., "Identification of transcriptional regulators in the mouse immune system," *Nature Immunology*, vol. 14, no. 6, pp. 633–643, 2013.

[12] S. Razick, G. Magklaras, and I. M. Donaldson, "iRefIndex: a consolidated protein interaction database with provenance," *BMC Bioinformatics*, vol. 9, article 405, 2008.

[13] B. Turner, S. Razick, A. L. Turinsky et al., "iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence," *Database*, vol. 2010, Article ID baq023, 2010.

[14] A. L. Turinsky, S. Razick, B. Turner, I. M. Donaldson, and S. J. Wodak, "Navigating the global protein-protein interaction landscape using iRefWeb," *Methods in Molecular Biology*, vol. 1091, pp. 315–331, 2014.

[15] M. P. H. Stumpf, T. Thorne, E. De Silva et al., "Estimating the size of the human interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 19, pp. 6959–6964, 2008.

[16] J. I. Fuxman Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, and A. J. Walhout, "Using networks to measure similarity between genes: association index selection," *Nature Methods*, vol. 10, no. 12, pp. 1169–1176, 2013.

[17] A. L. Barabási, "Scale-free networks: a decade and beyond," *Science*, vol. 325, no. 5939, pp. 412–413, 2009.

[18] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[19] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8577–8582, 2006.

[20] T. Clancy, E. A. Rødland, S. Nygard, and E. Hovig, "Predicting physical interactions between protein complexes," *Molecular and Cellular Proteomics*, vol. 12, no. 6, pp. 1723–1734, 2013.

[21] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[22] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[23] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[24] U. Bodenhofer, A. Kothmeier, and S. Hochreiter, "APCluster: an R package for affinity propagation clustering," *Bioinformatics*, vol. 27, no. 17, pp. 2463–2464, 2011.

[25] M. J. Brusco and H. Köhn, "Comment on "clustering by passing messages between data points"," *Science*, vol. 319, no. 5864, p. 726, 2008.

[26] B. J. Frey and D. Dueck, "Response to comment on "clustering by passing messages between data points"," *Science*, vol. 319, no. 5864, p. 726, 2008.

[27] M. Mézard, "Where are the exemplars?" *Science*, vol. 315, no. 5814, pp. 949–951, 2007.

[28] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[29] B. T. Sherman, D. W. Huang, Q. Tan et al., "DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis," *BMC Bioinformatics*, vol. 8, article 426, 2007.

[30] S. Kirov, R. Ji, J. Wang, and B. Zhang, "Functional annotation of differentially regulated gene set using WebGestalt: a gene set predictive of response to ipilimumab in tumor biopsies," *Methods in Molecular Biology*, vol. 1101, pp. 31–42, 2014.

[31] J. Wang, D. Duncan, Z. Shi, and B. Zhang, "WEB-based GEne SeT analysis toolkit (WebGestalt): update 2013," *Nucleic Acids Research*, vol. 41, pp. W77–W83, 2013.

[32] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software Environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[33] J. Staerk and S. N. Constantinescu, "The JAK-STAT pathway and hematopoietic stem cells from the JAK2 V617F perspective," *JAK-STAT*, vol. 1, no. 3, pp. 184–190, 2012.

[34] M. M. Davis, "A prescription for human immunology," *Immunity*, vol. 29, no. 6, pp. 835–838, 2008.

[35] T. Clancy, M. Pedicini, F. Castiglione et al., "Immunological network signatures of cancer progression and survival," *BMC Medical Genomics*, vol. 4, article 28, 2011.

[36] K. B. Lorvik, O. A. Haabeth, T. Clancy, B. Bogen, and A. Corthay, "Molecular profiling of tumor-specific T1 cells activated in vivo," *Oncoimmunology*, vol. 2, no. 5, Article ID e24383, 2013.

[37] T. Shay, V. Jojic, O. Zuk et al., "Conservation and divergence in the transcriptional programs of the human and mouse immune systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 8, pp. 2946–2951, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 791238, 9 pages
http://dx.doi.org/10.1155/2014/791238

Review Article

Alzheimer's Disease and HLA-A2: Linking Neurodegenerative to Immune Processes through an In Silico Approach

Ricardo A. Cifuentes and Juan Murillo-Rojas

Area of Basic Sciences, Faculty of Medicine, Universidad Militar Nueva Granada, Bogotá, Colombia

Correspondence should be addressed to Ricardo A. Cifuentes; ricardo.cifuentesgarcia@gmail.com

Received 22 January 2014; Accepted 8 July 2014; Published 17 August 2014

Academic Editor: Francesco Pappalardo

There is a controversial relationship between HLA-A2 and Alzheimer's disease (AD). It has been suggested a modifier effect on the risk that depends on genetic loadings. Thus, the aims of this study were to evaluate this relationship and to reveal genes associated with both concepts the HLA-A gene and AD. Consequently, we did first a classical systematic review and a meta-analysis of case-control studies. Next, by means of an in silico approach, we used experimental knowledge of protein-protein interactions to evaluate the top ranked genes shared by both concepts, previously found through text mining. The meta-analysis did not show a significant pooled OR (1.11, 95% CI: 0.98 to 1.24 in Caucasians), in spite of the fact that four of the included studies had a significant OR > 1 and none of them a significant OR < 1. In contrast, the in silico approach retrieved nonrandomly shared genes by both concepts ($P = 0.02$), which additionally encode truly interacting proteins. The network of proteins encoded by *APP, ICAM-1, ITGB2, ITGAL, SELP, SELL, IL2, IL1B, CD4*, and *CD8A* linked immune to neurodegenerative processes and highlighted the potential roles in AD pathogenesis of endothelial regulation, infectious diseases, specific antigen presentation, and HLA-A2 in maintaining synapses.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative process of the central nervous system (CNS) that is clinically characterized by an impairment of memory and other cognitive functions [1]. It is recognized as a multifactorial illness with both genetic and nongenetic causes. There have been identified four major genes associated with inherited risk for AD: *presenilin-1*, *presenilin-2*, *amyloid precursor protein* (*APP*), and *apolipoprotein E*. Mutations in these genes cause dysregulation of amyloid precursor protein processing, and in particular of the handling of a proteolytic derivative termed *beta-amyloid* (Abeta) that ultimately causes neuronal dysfunction and death [2].

Some findings also suggest an immune involvement in AD. Telomere length of T cells has been inversely correlated with cognitive performance impairment, apoptosis, serum levels of TNF-α, and the proportion of $CD8^+$ T cells lacking expression of the CD28 costimulatory molecule [3]. There are augmented levels of $CD8^+$ T memory cells, down regulation of CD8 receptors, and increased reactivity of $CD4^+$ and $CD8^+$ T-cells [4]. With regard to disease stages, there have been reported alterations in subsets of $CD4^+$ cells in patients with mild AD, with decreased percentages of naive cells, elevated memory cells, and increased proportions of $CD4^+$ cells lacking CD28. T potentially regulatory cells, $CD4^+CD25^{+high}$, with a naive phenotype are also reduced in AD patients [5]. It has been observed in patients of severe stage that there is a significant TNF-α increase in serum as well as a significant decrease in $CD4^+$ lymphocytes [6].

Epidemiological data suggest that some determinants of AD might reside in genes from the human leukocyte antigen (HLA) that regulate immune inflammatory responses [7]. It has been described the association of AD with both HLA-B7 and HLA-A2 [1]. Some authors have also found increased frequency of HLA-A$*$01 and HLA-DRB1$*$03 alleles and decreased frequency of HLA-DRB1$*$09 in late-onset AD cases [8, 9]. But these associations have shown no consistency among different ethnic groups [1]; nearly every positive result has been followed by several studies that have failed

to replicate it or that have contradicted it [7]. In the case of HLA-DRB1∗03 and its linked TNF-α 2-1-2 haplotype (-308/A, -238/G and TNF-a2 polymorphisms), it has been described a protective effect against AD [10], contrary to the effect of the HLA-DRB1∗03 allele described above. Even more, some researchers have indicated that there is no compelling evidence of a strong, direct association between AD and any HLA class I or II allele [11]. Consequently, it has been suggested that there is a modifier effect on the risk that depends on genetic loadings and further analysis, considering both HLA and non-HLA genes, are therefore necessary [7, 10].

However, there is accumulated evidence that suggests the involvement of the *HLA-A* gene in the pathogenesis of AD. A meta-analysis of all studies available until the 2000th year supported previous evidence of an excess of HLA-A2 in AD [12]. More recently, it has been observed that HLA-A2 and APOE4 independently reduced the age at onset of AD through an effect that seems to be additive in a population from China [13] and that A2 homozygotes had an earlier onset of AD in a population from North-America [14]. With this panorama, the aims of this study were to evaluate the current evidence of the association between HLA-A2 and AD and reveal genes that can influence the relationship between HLA-A and AD, thus assisting to point out pathogenic pathways related to AD. Our analysis was made by means of a meta-analysis of case-control studies that evaluated this association, and by using experimental knowledge of protein-protein interactions to evaluate the top ranked genes that were shared by the concepts HLA-A and AD, which had previously been found through a text mining approach of the biomedical literature.

2. Materials and Methods

2.1. Search Strategy and Selection Criteria. A systematic review of electronic databases (PubMed, EMBASE) was done independently by two researchers. The final date for inclusion was June, 2013. The search strategy used MeSH terms and text words: "Alzheimer disease," "Alzheimer's disease," "Alzheimer," and "HLA." No other criteria were taken into account. The inclusion criteria were the following: (1) AD diagnosis established by using the *National Institute of Neurological and Communicative Disorders and Stroke* and *the Alzheimer's Disease and Related Disorders Association* (NINCDS-ADRDA), The *Consortium to Establish a Registry for Alzheimer's Disease* (CERAD) or the *Diagnostic and Statistical Manual* (DSM) criteria; (2) If AD diagnosis criteria mentioned in numeral 1 were not used, the article must mention that there was histopathological confirmation or that other causes of dementia were clinically excluded in the patients from the AD cohort; (3) indication in the title or in the abstract that a relationship between HLA and AD was evaluated; (4) case-control study design; (5) publication of sufficient original data on the HLA-A2 prevalence in cases and in controls to calculate reliable odds ratios (OR) [15]; (6) etiology of cases not related to the four major genes described in AD [2]; and (7) manuscript's publication in a peer-reviewed journal as a full paper.

2.2. Data Extraction and Meta-Analysis. The data collected from each study were as follows: first author and the year of study, country, the number of cases and controls typified as HLA-A2 or with the alternative classification of HLA-A. Calculations were done for each ethnic origin by using the Catmap package at R software [16] as previously described [15]. Briefly, OR were grouped by weighing individual OR by the inverse of their variance. Thus, the final effect OR and the 95% confidence interval (95%CI) were obtained by means of both random- and fixed-effects models. The fixed-effects model was used when the random-effects variance was less than or equal to zero and there was no heterogeneity, defined as $P < 0.10$ by the Cochran's (Q) test; otherwise, the random-effects model was chosen. Publication bias was evaluated by a sensitivity analysis.

2.3. Text Mining Approach. To find out the genetic similarity of the "HLA-A" and "Alzheimer's disease" concepts, we used the Anni software [17] because it uses the concept profile methodology that has proven to be effective in finding information in the form of associations in the biological domain [18], as previously described [19]. Briefly, we first mapped those concepts in the thesaurus of the software and built the concept profile for each one. These profiles corresponded to the weighted list made by all the genes mentioned in MedLine, so they were called genetic concept profiles (GCPs). To do this, we selected the 25.010 genes that belong to human beings from the thesaurus in Anni, and, then, we mined all the MedLine records that contained these genes in their text.

Next, we matched these two GCPs and analyzed the similarities between them. For this purpose, we obtained a cohesion score (CS) by using as an inclusive filter for matching the described 25.010 genes. To interpret the cohesion score we used a P value that gives the probability that the same CS or higher would be found in a random group of the same size. This P value was obtained by using the default parameter in Anni of 200 iterations. The contribution of each gene in the profile to the similarity between both GCPs was assessed in terms of percentage. The MedLine records that support a contribution higher than 0.1% to the similarity between GCPs were reviewed to verify that true genes, or the proteins they encode, were associated to the concepts "HLA-A" or "Alzheimer's disease." Associations with ambiguous terms were eliminated.

2.4. Evaluation of Shared Genes by a Protein-Protein Interaction Network. To analyze which of the proteins encoded by the genes with the highest contribution to the similarity between GCPs truly interact, a network analysis was done with the genes that contributed at least 0.1% to the CS. For this purpose, the software, Genes2networks, was employed because it provides a reference network of experimentally known protein-protein interactions [20]. Then, in order to find tightly connected proteins, the settings that were used to build the network were (1) no filter for minimum number of references, (2) the maximum links per reference were four, (3) a maximum pathway length of two, and (4) a significant

Table 1: Description of the included articles.

Study	Country/population	Cases/controls	Typing/technique	Diagnostic criteria	Reference
Henschke et al., 1978	Canada	34/239	Lymphotoxicity	Exclusion of other causes of dementia	[21]
Sulkava et al., 1980	Finland	32/35	Lymphotoxicity	Exclusion of other causes of dementia	[22]
Wilcox et al., 1980	United Kingdom	18/342	Lymphotoxicity	Exclusion of other causes of dementia	[23]
Whalley et al., 1980	United Kingdom	14/64	Lymphotoxicity	Exclusion of other causes of dementia	[24]
Majsky and Vojtechovsky, 1983	Czechoslovakia	38/301	Lymphotoxicity	Exclusion of other causes of dementia-Haschinsky	[25]
Reed et al., 1983	United States	44/100	Lymphotoxicity	Exclusion of other causes of dementia	[26]
Reisner et al., 1983	United States	52/305	Amos modified method	Histopathological confirmation	[27]
Renvoize, 1984	United Kingdom	124/458	Lympho-toxicity	Exclusion of other causes of dementia	[28]
Endo et al., 1986	Japan	122/66	Lympho-toxicity	DSM III	[36]
Small and Matsuyama, 1986	United States	36/25	Lympho-toxicity	DSM III	[29]
Payami et al., 1991	United States	54/263	Lymphotoxicity	NINCDS-ADRDA	[30]
Middleton et al., 1999	United Kingdom	95/45	PCR SSOP	Histopathological confirmation	[31]
Small et al., 1999	United States	479/233	PCR SSP	NINCDS-ADRDA	[32]
Harris et al., 2000	United Kingdom	178/161	PCR SSP	NINCDS-ADRDA	[12]
Lehmann et al., 2001	United Kingdom	55/73	PCR SSP	Histopathological confirmation CERAD	[11]
Araria-Goumidi et al., 2002	France	451/477	Duplex-PCR	DSM III and NINCDS-ADRDA	[33]
Listì et al., 2006	Italy	460/266	PCR SSP	NINCDS-ADRDA	[34]
Ma et al., 2008	China	160/167	SBT	NINCDS-ADRDA	[13]
Guerini et al., 2009	Italy	173/258	PCR SSP	NINCDS-ADRDA	[35]

AD: Alzheimer's disease, PCR SSOP: Polymerase chain reaction and sequence specific oligonucleotide probe, PCR SSP: Polymerase chain reaction and sequence specific primers, SBT: Sequence based typing, DSM: Diagnostic and Statistical Manual, NINCDS-ADRDA: National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association, CERAD: Consortium to Establish a Registry for Alzheimer's Disease.

z-score of 2.5 of the intermediate nodes, which was calculated through a binomial proportions test, as previously described [19].

3. Results

From Europe, North-America and Asia, nineteen studies with data from case-control studies (2619 cases and 3878 controls) fitted the selection criteria; detailed information on the 185 articles that were excluded is given in the supplementary Table S1 in the Supplementary Material available online at http://dx.doi.org/10.1155/2014/791238. 17 of the included studies were from Caucasians [11, 12, 21–35] and 2 from Asians [13, 36]; see Table 1. Regarding the meta-analysis, the HLA-A2 did not have a specific behavior of risk or of protection with a pooled OR of 1.11, 95%CI: 0.98 to 1.24 (QP value 0.02). None of the articles included in the meta-analysis showed a significant OR lesser than one, but in contrast four studies showed a significant OR higher than one. In the same line, exclusion of one study by means of the sensitivity analysis led in some cases to significant risk behavior of the HLA-A2 but never to a significant protective behavior (Figure 1).

As it was recognized the cross-reactivity of the HLA-A2 antigen with HLA-A28, and that sera containing antibodies against A2 and A28 supertypic determinants are frequently found [37], we did a meta-analysis in Caucasians with the 8 studies that only used molecular techniques [11–13, 31–35]; see Table 1. The model showed similar unspecific results with a pooled OR of 1.03, 95%CI: 0.9 to 1.19 (QP value 0.03). With regard to the studies from Asia, the model also showed a very unspecific pooled OR than can be found between 0.84 and 2.16 (QP value 0.19).

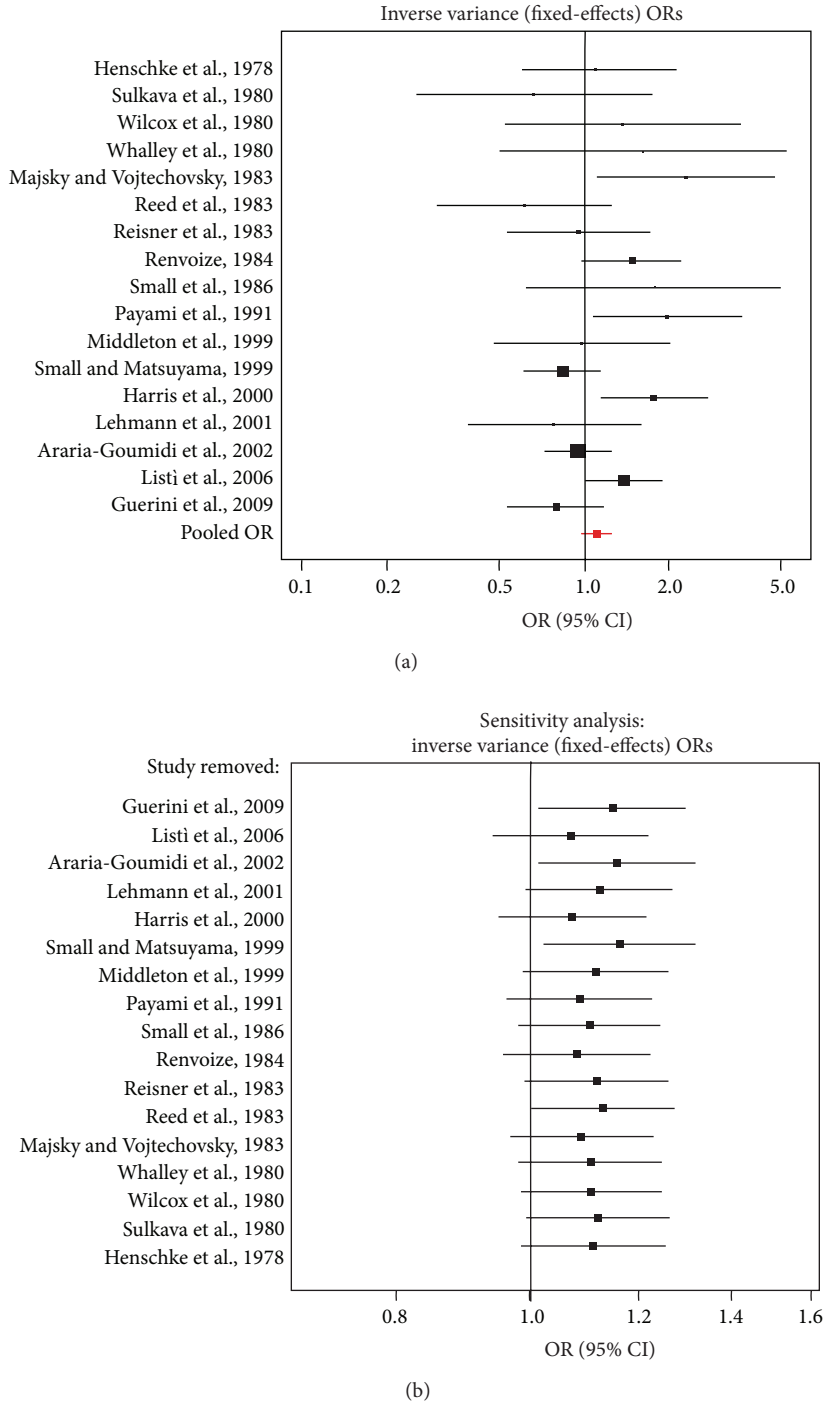


Figure 1: Forest plots of studies that relate HLA-A2 and AD. (a) Shows the effect summary OR (pooled OR) that takes into account all the studies. (b) Shows the pooled OR when each one of the studies was removed (sensitivity analysis).

Contrary to the nonconclusive results of the meta-analysis, the GCPs from both concepts AD and HLA-A were, not at random, genetically similar (CS P value 0.02). In addition to the *HLA-A*, 20 genes had a contribution higher than 0.1% to this similarity. As it was expected the weights of the genes involved in neuron remodeling and differentiation such as *APP* and *APLP2* were higher in the GCP of AD and the weights of the genes involved in immunity such as *HLA-DRB1* were higher in the GCP of HLA-A (Table 2).

Regarding the interaction analysis, proteins encoded from 10 of the 21 genes used as input were kept in the network, (Figure 2). Some genes shared by HLA-A and AD such as *APLP2*, *HLA-DRB1,* and *HFE* did not appear in the network despite their studied association with AD and/or HLA-A2

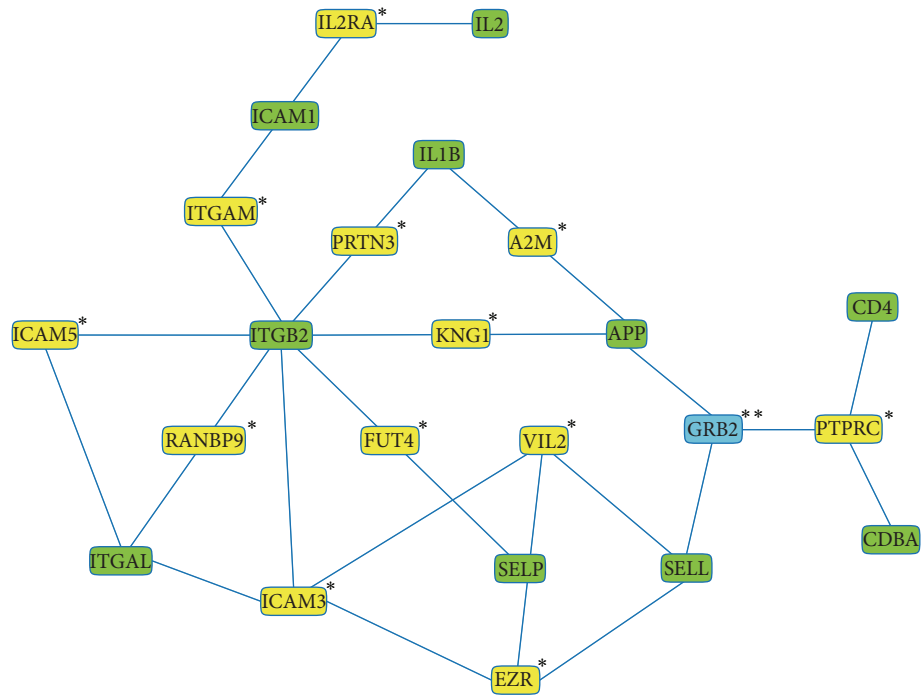


Figure 2: Interaction network of the proteins encoded by genes that contribute at least 0.1% to the cohesion score between HLA-A and AD. The nodes correspond to proteins encoded by the seed genes, to significant intermediary ones (indicated by one asterisk) and to a nonsignificant intermediary one (indicated by two asterisks).

Table 2: Genes with a contribution higher than 0.1% to the similarity between AD and HLA-A.

Gene	Percentage	Weight in AD	Weight in HLA-A
HLA-A	38.870	$1.257E-6$	$3.990E-2$
APLP2	21.170	$4.900E-3$	$5.546E-6$
APP	17.115	$7.000E-3$	$3.176E-6$
HLA-DRB1	5.576	$3.333E-6$	$2.200E-3$
CD8A	3.116	$1.434E-5$	$3.000E-4$
ICAM1	2.115	$5.694E-6$	$5.000E-4$
HLA-B	1.904	$1.067E-6$	$2.300E-3$
HLA-C	1.721	$1.472E-6$	$1.500E-3$
CD4	0.611	$2.812E-5$	$2.807E-5$
IFNG	0.391	$6.326E-6$	$7.995E-5$
HFE	0.223	$2.891E-6$	$9.967E-5$
IL1B	0.176	$2.309E-5$	$9.881E-6$
IL2	0.170	$8.815E-6$	$2.494E-5$
TYR	0.165	$2.515E-6$	$8.474E-5$
WT1	0.147	$1.599E-6$	$1.000E-4$
SELL	0.138	$1.748E-6$	$1.000E-4$
ITGB2	0.135	$4.105E-6$	$4.252E-5$
ITGAL	0.128	$2.750E-6$	$6.013E-5$
CD80	0.124	$4.628E-6$	$3.470E-5$
CD86	0.123	$3.332E-6$	$4.798E-5$
SELP	0.113	$2.683E-6$	$5.478E-5$

Cohesion score P value 0.02.

Table 3: Significance of intermediates sorted by z-score.

Node name	Links	Links in background	Links to seed	Links in subnetwork	z-score
FUT4	3	11429	2	29	22.86
ICAM5	4	11429	2	29	19.77
PRTN3	9	11429	2	29	13.10
IL2RA	22	11429	3	29	12.47
ICAM3	10	11429	2	29	12.41
VIL2	32	11429	3	29	10.25
ITGAM	15	11429	2	29	10.06
EZR	34	11429	3	29	9.93
KNG1	22	11429	2	29	8.23
RANBP9	22	11429	2	29	8.23
A2M	24	11429	2	29	7.86
PTPRC	35	11429	2	29	6.42
GRB2	196	11429	2	29	2.13

[8, 38, 39] and even the linkage disequilibrium with the HLA-A gene [40]. This could have been because of the strict threshold, a maximum pathway length of two, established to avoid weak interactions. Furthermore, the network had 13 intermediary nodes, 12 significant with a z-score above the cutoff of 2.5 (Table 3), thus indicating that the seed genes encode proteins that had strong and specific interactions.

In the graph, we found out two subnetworks (Figure 2): The first was made up of APP (involved in remodeling, differentiation, and apoptosis of neurons), ICAM1, ITGB2, ITGAL, SELP and SELL (involved in leukocyte adhesion, rolling over vascular surfaces, and transendotelial migration), and IL2 and IL1B (involved in T-cell proliferation, activation, and inflammation) and the second subnetwork was made up of CD4 and CD8A (involved in HLA classes I and II antigen presentation). These two subnetworks were connected by the growth factor receptor-bound protein (GRB2), the unique intermediate node that had a nonsignificant z-score due to low specificity because of its many links (Table 3), in other words, because it is a molecule involved in many cellular processes [41].

4. Discussion

Despite having new studies with big samples and homogeneous criteria for inclusion of AD patients (i.e., NINCDS-ADRDA) compared to the meta-analysis that associated HLA-A2 and AD more than ten years ago [12], we did not find conclusive results by this classical approach. The HLA-A2 showed to be a mild risk factor of AD with significant results only in some populations, thus suggesting that there are processes that influence this relationship. In contrast, the in silico approach retrieved nonrandomly shared genes by the concepts of HLA-A and AD ($P = 0.02$), that additionally encode truly interacting proteins. Proteins encoded by *APP*, *ICAM-1*, *ITGB2*, *ITGAL*, *SELP*, *SELL*, *IL2*, *IL1B*, *CD4,* and *CD8A* interact and were statistically and experimentally related to both concepts: HLA-A and AD. The network of interacting proteins highlighted specific processes, thus assisting to point out relevant pathogenic pathways that linked immunity to AD. Immune processes such as leukocyte adhesion and transendothelial migration, peptide presentation, and T-cell activation and proliferation were linked to processes traditionally involved in the pathogenesis of AD such as remodeling and apoptosis of neurons.

The results of our meta-analysis point out the importance of finding out relevant gene networks than can influence the relationship between HLA-A and AD and are in the same line with a previous analysis of the HLA-B, another gene of the same complex genomic region. In Oxford, researchers from UK confirmed in 2006 the association between HLA-B7 and AD, which was previously found in other people from the same city in 2001. However, this association was not found in populations from Cambridge, UK, and Montreal, Canada in spite the fact that all were of Caucasian origin. That is why, it was suggested a geographical specificity that could be due to different interactions with other processes of environmental, genetic or epigenetic origin [42].

Regarding the highlighted processes by our analysis of gene networks, it has been observed alterations of endothelial regulation in AD. IL1A, IL1B, IL2, IL8, IFNγ, and TNFα have been found to be associated with senile plaques. Some of them, IL8 and IFNγ were also significantly increased in plasma. Abnormal secretion of cytokines due to immune activation may impair the regulation of endothelial cells and induce altered pathways of adhesion molecules. There have been observed lower levels of P-selectin and L-selectin in AD and lowest in patients with the highest cognitive decline, thus leading to impaired regulation of both endothelial function and leukocyte migration [43]. With this landscape, infections at the level of the vasculature may be a key initiating factor in the pathogenesis of neurodegenerative diseases such as sporadic AD. Some observations have shown that *C. pneumoniae* infection stimulates transendothelial entry of monocytes through human brain endothelial cells (HBMECs). This entry

is facilitated by the upregulation of VCAM-1 and ICAM-1 on HBMECs and a corresponding increase of ITGB2-ITGAL (LFA-1), VLA-4, and ITGB2-ITGAM (MAC-1) on monocytes [44].

Another important process is the antigen presentation as it has been demonstrated the highly immunogenic properties of one specific HLA class II allele in a model of AD. It was observed that Abeta was effectively cleared from the brain parenchyma and brain microglial activation was reduced in long-term therapeutic immunization of an AD mouse model bearing the DRB1∗1501 allele [45]. Regarding to HLA-A2, it is a HLA class I protein that not only plays roles in the initiation of antigen receptor signaling, but is also expressed in neurons throughout the CNS. Neuronal HLA class I is upregulated after exposure to cytokines and functions as a mediator of synaptic plasticity during development of the visual system. Additional studies suggest that HLA class I may regulate the ability of neurons to maintain synapses. HLA class I mediated signaling has been studied, and it was observed that specifically HLA-A2 is substrate for the Alzheimer's disease-associated presenilin-1/gamma-secretase [46].

All in all, there is important evidence of the association between the described processes. A Genome Wide Association study found a significant single nucleotide polymorphism associated with AD within *GAB2*, which encodes "GRB2-associated binding protein 2" (Gab2). Gab2 binds GRB2 that also binds tau, APP, presenilin-1, and presenilin-2. Consequently, Gab2 could conceivably modulate APP processing and/or tau phosphorylation via its interaction with GRB2 [47]. Additionally, GRB2 is a known adapter with a recently described role in antigen receptor signaling as well as lymphocyte development [41].

5. Conclusion

Our review gives support to the immune involvement in AD. However, we not only find out a network of interacting proteins that links neurodegenerative to immune processes but that also gives hints for further research such as infectious diseases that alter the endothelial regulation as possible starting factor in AD, the role of GRB2 as a molecule that links antigen presentation with neuronal processes or the HLA-A2 role in the typical synaptic loss of AD. Thus, taking into account the described findings and the current overwhelming amount of data it seems highly advisable to combine in silico techniques with classical approaches such as systematic reviews or meta-analyses to find useful information.

Abbreviations

95%CI:	95% confidence interval
Abeta:	Beta-amyloid
AD:	Alzheimer disease
APP:	Amyloid precursor protein
CERAD:	The Consortium to Establish a Registry for Alzheimer's Disease
CNS:	Central nervous system
CS:	Cohesion score
DSM:	Diagnostic and Statistical Manual
GCPs:	Genetic concept profiles
HBMECs:	Human brain endothelial cells
HLA:	Human leukocyte antigen
NINCDS-ADRDA:	National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association
OR:	Odds ratio.

Conflict of Interests

The authors declare no conflict of interests.

Authors' Contribution

R. A. Cifuentes performed the study design, data extraction, meta-analysis, text mining, network analysis, and redaction of the paper and J. Murillo-Rojas participated in the study design, data extraction, and supplementary tables.

Acknowledgment

This work was supported by the Universidad Militar Nueva Granada, Bogotá, Colombia.

References

[1] C. Alves, S. Veiga, T. Souza, M. B. Toralles, and A. L. Da Silva-Bacellar, "The role of the human histocompatibility antigens in the pathogenesis of neurological disorders," *Revista de Neurologia*, vol. 44, no. 5, pp. 298–302, 2007.

[2] P. H. St George-Hyslop and A. Petit, "Molecular biology and genetics of Alzheimer's disease," *Comptes Rendus Biologies*, vol. 328, no. 2, pp. 119–130, 2005.

[3] L. A. Panossian, V. R. Porter, H. F. Valenzuela et al., "Telomere shortening in T cells correlates with Alzheimer's disease status," *Neurobiology of Aging*, vol. 24, no. 1, pp. 77–84, 2003.

[4] K. Schindowski, A. Eckert, J. Peters et al., "Increased T-cell reactivity and elevated levels of CD8+ memory T-cells in Alzheimer's disease-patients and T-cell hyporeactivity in an Alzheimer's disease-mouse model: Implications for immunotherapy," *NeuroMolecular Medicine*, vol. 9, no. 4, pp. 340–354, 2007.

[5] A. Larbi, G. Pawelec, J. M. Witkowski et al., "Dramatic shifts in circulating CD4 but not CD8 T cell subsets in mild Alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 17, no. 1, pp. 91–103, 2009.

[6] K. Bonotis, E. Krikki, V. Holeva, C. Aggouridaki, V. Costa, and S. Baloyannis, "Systemic immune aberrations in Alzheimer's disease patients," *Journal of Neuroimmunology*, vol. 193, no. 1-2, pp. 183–187, 2008.

[7] G. Candore, C. R. Balistreri, G. Colonna-Romano, D. Lio, and C. Caruso, "Major histocompatibility complex and sporadic Alzheimer's disease: a critical reappraisal," *Experimental Gerontology*, vol. 39, no. 4, pp. 645–652, 2004.

[8] D. Neill, M. D. Curran, D. Middleton et al., "Risk for Alzheimer's disease in older late-onset cases is associated with HLA-DRB1∗03," *Neuroscience Letters*, vol. 275, no. 2, pp. 137–140, 1999.

[9] F. R. Guerini, C. Tinelli, E. Calabrese et al., "HLA-A* 01 is associated with late onset of Alzheimer's disease in Italian patients," *International Journal of Immunopathology and Pharmacology*, vol. 22, no. 4, pp. 991–999, 2009.

[10] D. Culpan, S. H. MacGowan, J. M. Ford et al., "Tumour necrosis factor-α gene polymorphisms and Alzheimer's disease," *Neuroscience Letters*, vol. 350, no. 1, pp. 61–65, 2003.

[11] D. J. Lehmann, H. Wiebusch, S. E. Marshall et al., "HLA class I, II & III genes in confirmed late-onset Alzheimer's disease," *Neurobiology of Aging*, vol. 22, no. 1, pp. 71–77, 2001.

[12] J. M. Harris, A. M. Cumming, N. Craddock, D. St Clair, and C. L. Lendon, "Human leucocyte antigen-A2 increases risk of Alzheimer's disease but does not affect age of onset in a Scottish population," *Neuroscience Letters*, vol. 294, no. 1, pp. 37–40, 2000.

[13] S. L. Ma, N. L. S. Tang, C. W. C. Tam et al., "Association between HLA-A alleles and Alzheimer's disease in a Southern Chinese community," *Dementia and Geriatric Cognitive Disorders*, vol. 26, no. 5, pp. 391–397, 2008.

[14] S. Zareparsi, D. M. James, J. A. Kaye, T. D. Bird, G. D. Schellenberg, and H. Payami, "HLA-A2 homozygosity but not heterozygosity is associated with Alzheimer disease," *Neurology*, vol. 58, no. 6, pp. 973–975, 2002.

[15] R. A. Cifuentes, A. Rojas-Villarraga, and J.-M. Anaya, "Human leukocyte antigen class II and type 1 diabetes in Latin America: a combined meta-analysis of association and family-based studies," *Human Immunology*, vol. 72, no. 7, pp. 581–586, 2011.

[16] K. K. Nicodemus, "Catmap: case-control and TDT meta-analysis package," *BMC Bioinformatics*, vol. 9, article 130, 2008.

[17] R. Jelier, M. J. Schuemie, A. Veldhoven, L. C. J. Dorssers, G. Jenster, and J. A. Kors, "Anni 2.0: a multipurpose text-mining tool for the life sciences," *Genome Biology*, vol. 9, no. 6, article R96, 2008.

[18] R. Jelier, G. Jenster, L. C. J. Dorssers et al., "Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation," *BMC Bioinformatics*, vol. 8, article 14, 2007.

[19] R. A. Cifuentes, D. Restrepo-Montoya, and J. Anaya, "The autoimmune tautology: an in silico approach," *Autoimmune Diseases*, vol. 2012, Article ID 792106, 2012.

[20] S. I. Berger, J. M. Posner, and A. Ma'ayan, "Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases," *BMC Bioinformatics*, vol. 8, article 372, 2007.

[21] P. J. Henschke, D. A. Bell, and R. D. T. Cape, "Alzheimer's disease and HLA," *Tissue Antigens*, vol. 12, no. 2, pp. 132–135, 1978.

[22] R. Sulkava, S. Koskimies, J. Wikstrom, and J. Palo, "HLA antigens in Alzheimers disease," *Tissue Antigens*, vol. 16, no. 2, pp. 191–194, 1980.

[23] C. B. Wilcox, E. A. Caspary, and P. O. Behan, "Histocompatibility antigens in Alzheimer's disease. A preliminary study," *European Neurology*, vol. 19, no. 4, pp. 262–265, 1980.

[24] L. J. Whalley, S. J. Urbaniak, C. Darg, J. F. Peutherer, and J. E. Christie, "Histocompatibility antigens and antibodies to viral and other antigens in Alzheimer pre-senile dementia," *Acta Psychiatrica Scandinavica*, vol. 61, no. 1, pp. 1–7, 1980.

[25] A. Majsky and M. Vojtechovsky, "Non-relationship between the HLA system and the senile form of Alzheimer's disease," *Archiv fur Psychiatrie und Nervenkrankheiten*, vol. 233, no. 5, pp. 381–383, 1983.

[26] E. Reed, D. Thompson, R. Mayeaux, and N. Suciu Foca, "HLA antigens in Alzheimer's disease," *Tissue Antigens*, vol. 21, no. 2, pp. 164–167, 1983.

[27] E. Reisner, A. Heyman, T. Weinberg, D. Dawson, and E. Ciftan, "Lack of association between Alzheimer's disease and histocompatibility antigens," *Tissue Antigens*, vol. 21, no. 1, pp. 31–34, 1983.

[28] E. B. Renvoize, "An HLA and family study of Alzheimer's disease," *Psychological Medicine*, vol. 14, no. 3, pp. 515–520, 1984.

[29] G. W. Small and S. S. Matsuyama, "HLA-A2 as a possible marker for early-onset Alzheimer disease in men," *Neurobiology of Aging*, vol. 7, no. 3, pp. 211–214, 1986.

[30] H. Payami, J. Kaye, W. Becker, D. Norman, and P. Wetzsteon, "HLA-A2, or a closely linked gene, confers susceptibility to early-onset sporadic Alzheimer's disease in men," *Neurology*, vol. 41, no. 10, pp. 1544–1548, 1991.

[31] D. Middleton, H. Mawhinney, M. D. Curran et al., "Frequency of HLA-A and B alleles in early and late-onset Alzheimer's disease," *Neuroscience Letters*, vol. 262, no. 2, pp. 140–142, 1999.

[32] G. W. Small, W. K. Scott, S. Komo et al., "No association between the HLA-A2 allele and Alzheimer disease," *Neurogenetics*, vol. 2, no. 3, pp. 177–182, 1999.

[33] L. Araria-Goumidi, J. C. Lambert, D. Cottel, P. Amouyel, and M. C. Chartier-Harlin, "No association of the HLA-A2 allele with Alzheimer's disease," *Neuroscience Letters*, vol. 335, no. 2, pp. 75–78, 2002.

[34] F. Listì, G. Candore, C. R. Balistreri et al., "Association between the HLA-A2 allele and Alzheimer disease," *Rejuvenation Research*, vol. 9, no. 1, pp. 99–101, 2006.

[35] F. R. Guerini, E. Calabrese, C. Agliardi et al., "Association study of the HLA-A2 allele in Italian Alzheimer disease patients," *Neurobiology of Aging*, vol. 30, no. 12, pp. 2082–2083, 2009.

[36] H. Endo, T. Yamamoto, and F. Kuzuya, "HLA system in senile dementia of Alzheimer type and multi-infarct dementia in Japan," *Archives of Gerontology and Geriatrics*, vol. 5, no. 1, pp. 51–56, 1986.

[37] M. A. Fernandez-Viña, M. Falco, Y. Sun, and P. Stastny, "DNA typing for HLA class I alleles: I. Subsets of HLA-A2 and of -A28," *Human Immunology*, vol. 33, no. 3, pp. 163–173, 1992.

[38] A. Tuli, M. Sharma, X. Wang et al., "Amyloid precursor-like protein 2 association with HLA class i molecules," *Cancer Immunology, Immunotherapy*, vol. 58, no. 9, pp. 1419–1431, 2009.

[39] A. P. Correia, J. P. Pinto, V. Dias, C. Mascarenhas, S. Almeida, and G. Porto, "CAT53 and HFE alleles in Alzheimer's disease: a putative protective role of the C282Y HFE mutation," *Neuroscience Letters*, vol. 457, no. 3, pp. 129–132, 2009.

[40] J. Vieira, C. S. Cardoso, J. Pinto et al., "A putative gene located at the MHC class I region around the D6S105 marker contributes to the setting of $CD8^+$ T-lymphocyte numbers in humans," *International Journal of Immunogenetics*, vol. 34, no. 5, pp. 359–367, 2007.

[41] I. K. Jang, J. Zhang, and H. Gu, "Grb2, a simple adapter with complex roles in lymphocyte development, function, and signaling," *Immunological Reviews*, vol. 232, no. 1, pp. 150–159, 2009.

[42] D. J. Lehmann, M. C. N. M. Barnardo, S. Fuggle et al., "Replication of the association of HLA-B7 with Alzheimer's disease: a role for homozygosity?" *Journal of Neuroinflammation*, vol. 3, article 33, 2006.

[43] M. M. Corsi, F. Licastro, E. Porcellini et al., "Reduced plasma levels of P-selectin and L-selectin in a pilot study from Alzheimer disease: relationship with neuro-degeneration," *Biogerontology*, vol. 12, no. 5, pp. 451–454, 2011.

[44] A. MacIntyre, R. Abramov, C. J. Hammond et al., "Chlamydia pneumoniae infection promotes the transmigration of monocytes through human brain endothelial cells," *Journal of Neuroscience Research*, vol. 71, no. 5, pp. 740–750, 2003.

[45] V. Zota, A. Nemirovsky, R. Baron et al., "HLA-DR alleles in amyloid β-peptide autoimmunity: a highly immunogenic role for the DRB1$*$1501 allele," *Journal of Immunology*, vol. 183, no. 5, pp. 3522–3530, 2009.

[46] B. W. Carey, D. Y. Kim, and D. M. Kovacs, "Presenilin/γ-secretase and α-secretase-like peptidases cleave human MHC Class I proteins," *Biochemical Journal*, vol. 401, no. 1, pp. 121–127, 2007.

[47] L. Bertram and R. E. Tanzi, "Genome-wide association studies in Alzheimer's disease," *Human Molecular Genetics*, vol. 18, no. 2, pp. R137–R145, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 410457, 19 pages
http://dx.doi.org/10.1155/2014/410457

Research Article

On the Coupling of Two Models of the Human Immune Response to an Antigen

Bárbara de M. Quintela, Rodrigo Weber dos Santos, and Marcelo Lobosco

Laboratory of Computational Physiology and High-Performance Computing (FISIOCOMP), Graduate Program in Computational Modeling, UFJF, Rua José Lourenço Kelmer s/n, Campus Universitário, Bairro São Pedro, 36036-900 Juiz de Fora, MG, Brazil

Correspondence should be addressed to Bárbara de M. Quintela; barbara@ice.ufjf.br

Received 31 January 2014; Revised 15 April 2014; Accepted 15 April 2014; Published 22 July 2014

Academic Editor: Francesco Pappalardo

The development of mathematical models of the immune response allows a better understanding of the multifaceted mechanisms of the defense system. The main purpose of this work is to present a scheme for coupling distinct models of different scales and aspects of the immune system. As an example, we propose a new model where the local tissue inflammation processes are simulated with partial differential equations (PDEs) whereas a system of ordinary differential equations (ODEs) is used as a model for the systemic response. The simulation of distinct scenarios allows the analysis of the dynamics of various immune cells in the presence of an antigen. Preliminary results of this approach with a sensitivity analysis of the coupled model are shown but further validation is still required.

1. Introduction

Systems biology is an emerging interdisciplinary area of science that advocates a distinct perspective on the study of biological phenomena, particularly focusing on understanding a system's structure and its dynamics [1]. The systems biology approach often involves the use of mathematical and computational techniques in the development of mechanistic models that describes complex interactions in biological systems.

One complex biological system that can benefit from the systems biology approach is the immune system (IS). The IS is composed by a large number of cells, molecules, tissues, and organs that form a complex network that interact with each other in order to protect the body against pathogenic agents [2]. The IS of vertebrates is composed by three layers of defense: (a) the physical barriers; (b) innate IS; and (c) the adaptive IS.

The physical barriers are composed by the skin and mucous membranes that form a shield against the pathogenic agents. If this shield is crossed by pathogens, they encounter cells and molecules of the innate IS, such as proteins of the complement system and macrophages, that immediately develop a response to them. The macrophages phagocyte the pathogens and produce proteins called cytokines that signal to other innate cells that their help is needed. Recruitment of more innate cells is essential for effective control of infections [3]. Some of the innate cells such as the macrophages and dendritic cells act as antigen presenting cells (APCs), playing a pivotal role on activating the third layer of defense, which is the adaptive IS response. B and T lymphocytes are some of the main cells of this third layer response. The presence of these cells is extremely important because they can adapt to deal with almost any invader. B cells in its plasma form secrete antibodies. Antibodies bind to pathogens, which turns the latter more susceptible to phagocytosis (in a process called opsonization). Three main types of T cells are known: (a) killer T cells (also known as cytotoxic lymphocytes), (b) helper T cells, and (c) regulatory T cells. Killer T cells induce infected cells to commit suicide, in a process called apoptosis;

helper T (Th2) cells produce cytokines and help priming B cells; and regulatory T cells act in the regulation of the response, although the complete process is still unclear [2].

A large number of works have proposed models to describe the IS [4–13]. A great introduction to previous models of the IS is available in the work of Perelson and Weisbuch [14] and more recently the state of the art on representing the IS was presented by Narang et al. [15].

Computational models of the adaptive IS are very often developed using pure mathematical tools, such as ordinary differential equations (ODEs), to describe the behavior of its components and their relationships, although other tools, such as system dynamics [16, 17], cellular automata [18–24], agent-based systems [25–36], and complex adaptive systems [37], are also used. Some works focus only on modeling the innate IS [38–40], which is responsible for activating the adaptive response.

The use of more than one approach to model the immune response is not novel and there are different works that consider differential equations together with cellular automata [23], agent-based Systems [41], and system dynamics [16]. And also, the assumption of different scales of the immune system was already considered by Kirschner [33]. However, none of the previous approaches so far represented the immune response considering the spatial features such as, cellular movement, diffusion, and chemoattraction modeled with PDEs and the dynamics of antibody activation modeled with ODEs, which is the approach used in this work. Differential equations were chosen due to the advantage of dealing with distinct scales, in comparison with other approaches, and the possibility of numerical analysis of the model.

A previous work presented a mechanistic computational model of the innate immune response to a general pathogen [42]. This pathogen is represented by lipopolysaccharide (LPS) that is present in the outer membrane of Gram-negative bacteria. That model represents the behavior of the main defense cells, such as macrophages, and molecules, such as proinflammatory cytokines (TNF-α and IL-8) and anti-inflammatory cytokines (IL-10). A set of PDEs was used to reproduce important phenomena such as the temporal order of cells arriving at the local of infection, the production of proinflammatory and anti-inflammatory cytokines, and the chemotaxis phenomenon. The model has been extended (a) to allow the use of a three-dimensional domain in order to better represent the site of infection and (b) to use parallel programming techniques to guarantee a reasonable simulation time [43]. This work extends our previous models [42, 43], enabling the innate IS to activate the adaptive one. The main contribution of this work is a mathematical model that reproduces the dynamics of both innate and adaptive IS, coupling for this purpose models of different nature and scales. The adaptive model chosen to be coupled with the innate model consists of a set of equations, based on the mathematical model of pneumonia which is described in [44]. Besides, the coupled model represents a more complete scenario: the dynamics of the cells and molecules inside the tissue as well as the communication through lymph and blood vessels with the nearest lymph node. We must stress that the main contribution of this work is the new way used to couple models of distinct scales and aspects of the immune system, represented by ODEs and PDEs. In this way, both models used to represent the innate and the adaptive immune system could be replaced by other models with slight modifications. In fact, several recent models represent the dynamics of acute inflammatory response in the lung [45–49]. The model proposed by [44] was chosen due to the availability of all parameters needed to implement the model.

The interest in modeling spatial features of the immune response is due to the increasing availability of noninvasive imaging techniques mainly on the last decade [50]. Besides, the spatial information of an individual could be provided at any time to validate the *in silico* models. A few examples of recent works on noninvasive imaging in immunology are [51–55]. Thus, an important aspect considered in choosing pneumonia as an example to illustrate the new coupling technique and why to use PDEs is the fact that this disease causes damages in the tissue that can be observed by medical imaging techniques. The simulation of the coupled model can generate as a result an image that represents the damage caused by bacteria to the local tissue, in this example the lung (alveolar space).

Related works on coupling models of the IS deal with different scales to represent the trigger for innate response and activation of acquired response. At a molecular level [56] used coupled ODEs models to understand the behavior of the proteins involved in the process of antigenic presentation. At a cellular level there are works that couple models employing agent-based systems with ODEs [31, 33, 57], or with system dynamics [16], while others use only PDEs [39, 58], only ODEs [45, 59], or DDEs [60] to achieve this purpose. A similar approach using only ODEs to represent tuberculosis dynamics can be found in [61]. The model proposed in this work uses both PDEs and ODEs to describe the entire dynamics. This is, for the best of our knowledge, the first work that couples PDEs and ODEs to describe the dynamics of innate and adaptive IS into a piece of tissue, including the activation of the adaptive IS by the innate IS.

This work is organized as follows. Section 2 presents the biological model used in this work and the coupling of the mathematical models. The IS model presented in this work was simplified when compared to our previous models [42, 43]. The complete model was not used in order to focus on the integration of the two different models: (a) local tissue and (b) lymph nodes. Section 3 presents its computational implementation. The results obtained by the models, the discussion, and a sensitivity analysis of the coupled model are presented in Section 4. Finally, Section 5 draws our conclusions and present our future works.

2. Materials and Methods

2.1. Biological Model. According to [44], there are several microorganisms that could be etiological agents of destructive pneumonia. However, the inflammatory and destructive process in lung tissue cannot be started unless there is a malfunctioning of local and general defense mechanisms

[62]. The multiplication of microorganisms occurs between 10 and 14 days and causes the disease which, in case of recovery, ends as a result of destruction of bacteria by antibodies, macrophages, and neutrophils. Initially, we consider a simplified scenario where there are only resting macrophages located in the tissue (as an example we chose the lung tissue) and T- and B-lymphocytes in the lymph nodes. After the injection of an antigen (A), those resting macrophages (RM) that encounter an antigen become activated macrophages (AM) and start producing proinflammatory cytokines (PC) to recruit other immune cells to the site of infection. Activated macrophages act as antigen presenting cells (APCs) migrating to the closest lymph node through lymph vessels (Figure 1) and exhibiting the antigen to the lymphocytes which initiates the activation and differentiation of T-lymphocytes into T-helper 2 lymphocytes (Th2) and activation and differentiation of B-lymphocytes into plasma cells.

Plasma cells are mass producers of antibodies (Figure 2). Those antibodies head to the local of the infection in the lung tissue through blood vessels. As soon as they reach an antigen they do what is called opsonization of the antigen, the process by which an antigen is marked for ingestion and destruction by a phagocyte (Figure 3) [63].

2.2. Coupling of Models. Marchuk uses DDEs to model the processes involved in pneumonia [44]. In this work, some assumptions are made to model that scenario.

(i) We do not consider the delay on biological processes such as T lymphocyte clonal expansion and antibodies production and release.

(ii) We consider only the temporal behavior of the cells inside the lymph node and the spatiotemporal behavior of the cells in the lung tissue. Thus, we expect to visualize the damage caused to the lung parenchyma.

In order to represent the immune response, PDEs based on a previous model of the innate response [42] have been employed to model the spatial and temporal behavior of the following components:

S. aureus bacteria (A);

resting macrophages (M_R);

activated macrophages (M_A);

specific antibodies (F).

Moreover, ODEs represent the cascade activation of the lymphocytes leading to the production and release of antibodies in the lymph node (LN), which is a simplification of the model in [44]:

T-lymphocytes (T);

B-lymphocytes (B);

plasma cells (P);

antibodies (F^L).

Macrophages act like APCs, activating the adaptive IS and the production and release of antibodies. The presentation and the later presence of antibodies in the infection site are only possible due to the coupling of the two distinct models presented herein. We assumed that the communication between the alveolar tissue and the nearest LN is guaranteed by the presence of lymph and blood vessels.

The linkage between the local and the systemic response is achieved by representing the APCs and the antibodies in both models. There is a PDE equation to model the activated macrophage behavior while they are inside the tissue and an ODE to model their concentration inside the LN. The flux of activated macrophages from the tissue to the LN involves the numerical integration of the macrophages in the tissue. The same style of approximation was adopted to model the migration of antibodies from the LN to the tissue. The PDEs based model and the ODEs based model are shown in the next subsection.

It must be stressed that pneumonia was chosen as an example to illustrate the framework proposed in this paper to couple models of different nature and scales; other models could be chosen to represent different infection processes caused by other diseases.

The general framework for the coupling of models proposed herein could be summarized as follows.

(1) Selection of two distinct models of the immune response: one related to a local response and other related to a systemic one.

(2) Identification of common or related variables among the chosen models, for example, the variables representing cells that migrate from one location to another during the response, and therefore act as APCs. If no common or related variable is found, a relation must then be created by adapting the models to the coupling as stated in the next step.

(3) Implementation of necessary adjustments to the set of equations. This step could involve changes on the equations which must represent the flux between two different locations; for example, there must be a term to represent the flux of antibodies leaving the LN and another one representing the arrival of them in the tissue. Those terms do not necessarily exist on the original models that are being coupled. The units of measurement must also be considered within this step. If necessary, conversions should be made to the units of the parameters to keep the correctness of the coupled model.

(4) Simulate the coupled model and validate the obtained results.

We expect to gain insights on the immune system response to distinct pathogenic agents with the coupling of models. It follows an example of the coupling of models to represent the immune response to *S. aureus* causing pneumonia.

2.3. PDEs Model. All the PDEs are modeled considering homogeneous Neumann boundary conditions.

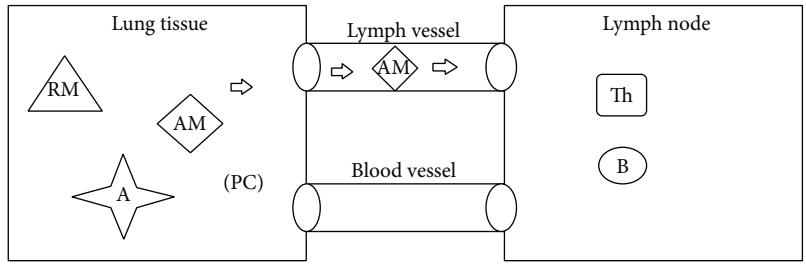


Figure 1: Communication between local tissue and lymph node; activation of resting macrophages (RM) and migration of activated macrophages (AM) to the lymph node.

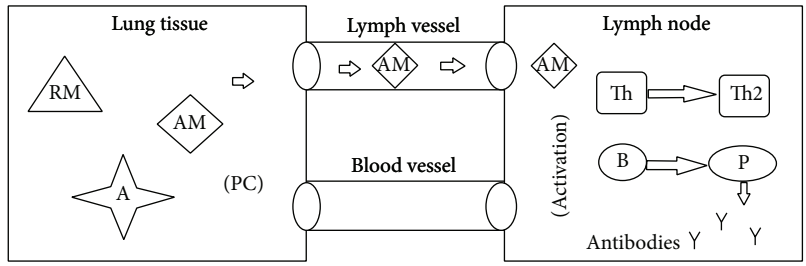


Figure 2: Activated macrophage (AM) stimulate lymphocytes by antigen presenting process; T cell differentiate in T-helper 2 (Th2) and B cell differentiate in plasma cell (P) which produces antibodies (Y).

2.3.1. S. aureus Bacteria (A). Equation (1) depicts the model for the *S. aureus* bacteria. The first term of (1) models the replication of bacteria at a rate β_A and carrying capacity is given by k_a. The second term gives its natural decay by the coefficient μ_A. Its engulfment by macrophages and other nonspecific defense cells is represented by the third and fourth terms of the same equation: λ_{MR} is the rate that accounts for the activation of macrophages and λ_{MA} is the destruction rate of bacteria by active macrophages.

$$\frac{\partial A}{\partial t} = \beta_A A\left(1 - \frac{A}{k_A}\right) - \mu_A A - \lambda_{MR} M_R A - \lambda_{MA} A M_A - \lambda_{AF|MR} A_F M_R - \lambda_{AF|MA} A_F M_A + D_A \Delta A, \quad (1)$$

$$A(x_0, y_0, z_0, 0) = A0, \qquad \frac{\partial A}{\partial t}(\cdot, t)\bigg|_{\partial\Omega} = 0.$$

The opsonization process for further phagocytosis in shown in the 5th and 6th terms; $\lambda_{AF|MR}$ represents the rate for destruction of opsonized bacteria by resting macrophages and $\lambda_{AF|MA}$ represents the rate for destruction by activated macrophages. The variable F is the amount of antibodies in the tissue modeled by (4) which is an important part of the coupling of models. The last term of this equation represents the diffusion of the bacteria in the lung tissue where D_A is the bacteria diffusion coefficient. Initially there is an injection of antigen only at one small part in the center of the tissue and it is assumed that there is no flux through the borders.

2.3.2. Macrophages (M_R, M_A). Macrophages are represented in two distinct states: resting (M_R) and activated (M_A) states. Initially, there are only resting macrophages in the tissue and they become activated after exposure to antigens (A). In the activated state, they play an important role in presenting and stimulating specific defense cells located in the LN.

Equation (2) represents the concentration of resting macrophages in the alveolar tissue, in which the first term accounts for their natural decay, the second term represents their activation, the third term represents the flux of resting macrophages entering the tissue from the blood, and the last one is the diffusion term. The natural decay rate is given by μ_{MR}, γ_{MA} is the rate in which resting macrophage becomes active and D_{M_R} is the resting macrophage diffusion coefficient. Consider

$$\frac{\partial M_R}{\partial t} = -\mu_{MR} M_R - \gamma_{MA} M_R A + \alpha_{M_R} \theta_{\mathrm{BV}}(x, y, z)\left(M_R^* - M_R\right) + D_{M_R} \Delta M_R, \quad (2)$$

$$M_R(x, y, z, 0) = M_{R0}, \qquad \frac{\partial M_R}{\partial t}(\cdot, t)\bigg|_{\partial\Omega} = 0.$$

The flux of macrophages entering the tissue depends on the localization of the blood vessels in the tissue. In order to represent that behavior α_{M_R} represents the rate of migration and θ_{BV} is a function that is equal to 1 where volumes are in contact with blood vessels and 0 otherwise.

Equation (3) represents the concentration of activated macrophages in the alveolar tissue after encountering antigens. Again, the first term models their natural decay, at a rate of μ_{MA}, the second term models their activation at a rate of

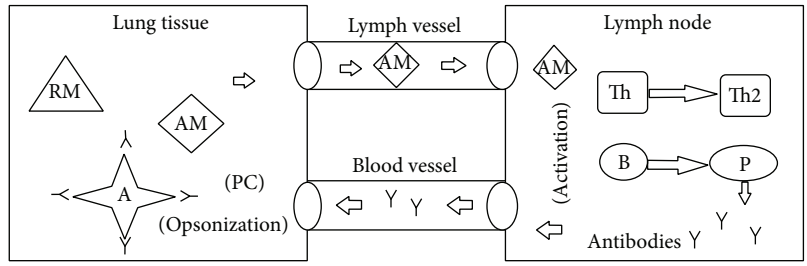


Figure 3: Antibodies (Y) migrate to the lung tissue through blood vessel and opsonize the antigen (A).

γ_{MA}, and the third one is the diffusion term, at a rate of D_{MA}. Consider

$$\frac{\partial M_A}{\partial t} = -\mu_{MA} M_A + \gamma_{MA} M_R A + D_{MA} \Delta M_A - \alpha_M \theta_{\mathrm{LV}}(x, y, z)\left(M_A - M_A^L\right), \quad (3)$$

$$M_A(x, y, z, 0) = MA_0, \qquad \left.\frac{\partial M_A}{\partial t}(\cdot, t)\right|_{\partial\Omega} = 0.$$

The last term of (3) models the connection between the activated macrophages in the tissue and the concentration that migrates to the nearest LN to act as APCs. It represents the flux of M_A between the local alveolar tissue and the LN through the lymph vessels. In this equation, M_A^L is the macrophage concentration in the LN, which dynamics is described by (5) and θ_{LV} is a function that is equal to 1 if the volumes are in contact with lymph vessels and 0 otherwise.

It is assumed that in the beginning of the simulation there are only resting macrophages over the tissue and there is no flux through the borders.

2.3.3. Antibodies (F). Equation (4) describes the antibody mechanics within the lung tissue. The first and second terms represent the antibodies consumption to defeat bacteria in the opsonization process, at rates depending on the state of the phagocyte cell: $\lambda_{FA|MR}$ for resting macrophages and $\lambda_{FA|MA}$ for activated macrophages. The third term models the diffusion process of antibodies in the tissue, at a rate of D_F, and the last term describes the flux of antibodies between the LN and the tissue, at a rate of α_F, in which F^L is the concentration of antibodies released by plasma cells in the LN (11). Consider

$$\frac{\partial F}{\partial t} = -\lambda_{FA|MR} F_A M_R - \lambda_{FA|MA} F_A M_A + D_F \Delta F + \alpha_F \theta_{\mathrm{BV}}(x, y, z)\left(F^L - F\right), \quad (4)$$

$$F(x_0, y_0, z_0, 0) = F0, \qquad \left.\frac{\partial F}{\partial t}(\cdot, t)\right|_{\partial\Omega} = 0.$$

The last term of (4) is part of the model coupling and was added to the PDEs model to make the connection between the antibodies released in the LN and their migration through the blood vessels to the infection site. It is assumed that there is no flux through the borders.

2.4. ODEs Model. In the ODEs model the cellular homeostasis is guaranteed by the addition of an equilibrium term in the equations corresponding to the adaptive IS. The idea is to preserve the minimum amount of adaptive IS cells in the body [44].

2.4.1. Macrophages (M_A^L). In order to perform the coupling, the antigen presentation needed to be represented as a trigger to the acquired response. The concentration of active macrophages inside the LN which migrated from the tissue was modeled by (5). This equation represents the active macrophages which migrated from the local alveolar tissue to the LN through the lymph vessels, at a rate of α_M

$$\frac{dM_A^L}{dt} = \alpha_M\left(M_A^T - M_A^L\right)\frac{V_{\mathrm{LV}}}{V_{\mathrm{LN}}}, \qquad M_A^L(0) = M_{A_0} \quad (5)$$

in which V_{LN} is the assumed volume of the LN and V_{LV} is the integral of the volumes where there is contact with lymph vessels given by

$$V_{\mathrm{LV}} = \int_\Omega \theta_{\mathrm{LV}}(x, y, z)\, d\Omega. \quad (6)$$

The average concentration of active macrophages in the tissue (M_A^T) was calculated by the integration of the values of active macrophages within the domain of simulation in contact with lymph vessels and is described by

$$M_A^T = \frac{1}{V_{\mathrm{LV}}}\int_\Omega \theta_{\mathrm{LV}}(x, y, z)\, M_A d\Omega, \quad (7)$$

where Ω represents the volume of the whole tissue simulated.

2.4.2. T-Lymphocytes (T). The T-helper lymphocytes are stimulated by active macrophages in the LN and play an important role in the activation of B-lymphocytes and plasma cells to start the production of specific antibodies against antigens. The first part of (8) represents the activation of Th2 cells, with its clonal expansion leading to the appearance of new cells. b_T is the rate for the stimulation of Th2 cells and ρ_T is the number of descendants Th2 cells created by single division. The second term represents the expenditure of Th2

cells to stimulate B cells, at a rate of b_p. Finally, the third term models the maintenance of the homeostasis in absence of antigenic stimulation

$$\frac{dT}{dt} = b_T\left(\rho_T T M_A^L - T M_A^L\right) - b_p M_A^L T B + \alpha_T\left(T^* - T\right), \quad T(0) = T_0 \tag{8}$$

in which T^* is the steady state value of the concentration of T-helper cells.

2.4.3. B-Lymphocytes (B). After B-lymphocytes cells have been stimulated by T cells and active macrophages in the LN, they start to proliferate and turn into plasma cells. Their proliferation is represented by the first term of (9), in which b_p^b is the rate for the stimulation of B cells and ρ_B is the number of new B cells as an outcome of the stimulation. Again, the second term shows the maintenance of the homeostasis in the absence of antigenic stimulation

$$\frac{dB}{dt} = b_p^b\left(\rho_B T M_A^L - T M_A^L B\right) + \alpha_B\left(B^* - B\right), \quad B(0) = B_0 \tag{9}$$

in which B^* is the steady state value for the concentration of B cells.

2.4.4. Plasma Cells (P). The plasma cells are generated from stimulated B cells, T-cells, and active macrophages in the LN and are the cells that release antibodies against the specific antigen that was presented by the APCs

$$\frac{dP}{dt} = b_p^p\left(\rho_P T M_A^L B\right) + \alpha_P\left(P^* - P\right), \quad P(0) = P_0. \tag{10}$$

The first term of (10) describes the generation and maturation of plasma cells from stimulated B cells, in which b_p^p is the rate for the stimulation of plasma cells and ρ_P is the number of new plasma cells. The last term is the maintenance of the homeostasis in which P^* is the steady state concentration of plasma cells.

2.4.5. Antibodies (F^L). Antibodies released by plasma cells in the LN are represented by

$$\frac{dF^L}{dt} = \rho_F P - \left(\alpha_F \theta_{\mathrm{BV}}(x,y,z)\left(F^L - F^T\right)\right)\frac{V_{\mathrm{BV}}}{V_{\mathrm{LN}}}, \quad F^L(0) = F_0 \tag{11}$$

in which the first term describes the production of antibodies, at rate ρ_F, by plasma cells and the last term represents the connection between the two models (PDEs and ODEs). The last term of (11) describes the flux of antibodies between the LN and the tissue, at a rate of α_F, on volumes with contact with the blood vessels given by θ_{BV}. F^T is the average number of antibodies in the tissue described by

$$F^T = \frac{1}{V_{\mathrm{BV}}}\int_\Omega \theta_{\mathrm{BV}}(x,y,z)\, F\, d\Omega, \tag{12}$$

where Ω is the tissue domain and V_{BV} is the integral of the volumes where there is contact with blood vessels given by

$$V_{\mathrm{BV}} = \int_\Omega \theta_{\mathrm{BV}}(x,y,z)\, d\Omega. \tag{13}$$

Table 1: Initial values of the coupled model.

Parameter	Value	Unit	Reference
A_0	2	Cell/mm^3	Estimated
M_{R_0}	4	Cell/mm^3	Estimated
M_{A_0}	0.0	—	[44]
F_0	0.0	—	[44]
T_0	0.0	—	[44]
B_0	0.0	—	[44]
P_0	0.0	—	[44]
F_0	0.0	—	[44]
T^*	$8.4 * 10^{-3}$	Cell/mm^3	[44]
B^*	$8.4 * 10^{-4}$	Cell/mm^3	[44]
P^*	$8.4 * 10^{-6}$	Cell/mm^3	[44]
F^*	0.0	Cell/mm^3	[44]
MR^*	4	Cell/mm^3	Estimated

2.5. Initial Conditions and Parameters. The initial conditions for the PDEs model which describe the process of formation of inflammatory site are shown in Table 1.

It was considered that initially only a small portion of the tissue had the presence of antigens and the domain of simulation was 10 mm^3. This initial injection of antigen was represented in the center of the hexahedral domain of simulation (between 3 mm and 7 mm over the axes). Initially it is also considered the presence of macrophages in its resting state equally distributed over the tissue.

Tables 2, 3, and 4 present the set of parameters used in the simulations. Almost all parameters used in the simulation were obtained in Marchuk [44] applying the necessary unit conversions and some fitting to the coupled model. The only exceptions are the diffusion coefficients D_{M_R}, D_{M_A}, and D_F which were estimated by Pigozzo et al. [42] and the diffusion coefficient of bacteria D_A based on the work of Haessler and Brown [62].

3. Implementation

The numerical method employed to solve the mathematical model was the finite difference method, a method commonly used in the numeric discretization of PDEs. The finite difference method is a method of resolution of differential equations that is based on the approximation of derivatives with finite differences [64, 65].

Table 2: Diffusion coefficients.

Parameter	Value	Unit	Reference
D_A	$3.7 * 10^{-5}$	mm^3/day	[62]
D_{MR}	$4.32 * 10^{-2}$	mm^3/day	[42]
D_{MA}	0.3	mm^3/day	[42]
D_F	$1.6 * 10^{-2}$	mm^3/day	[42]

Table 3: Replication, decay, activation, and phagocytosis rates.

Parameter	Value	Unit	Reference
β_A	2.0	1/day	[44]
k_A	50.0	cell/mm^3	Estimated
μ_A	0.1	1/day	[44]
μ_{MR}	0.033	1/day	[42]
μ_{MA}	0.07	1/day	[42]
γ_{AM}	$8.3 * 10^{-2}$	mm^3/cell∗day	[44]
λ_{MR}	$5.98 * 10^{-3}$	mm^3/cell∗day	[44]
λ_{MA}	$5.98 * 10^{-2}$	mm^3/cell∗day	[44]
$\lambda_{AF\|MR}$	$1.66 * 10^{-3}$	mm^6/$cell^2$∗day	[44]
$\lambda_{AF\|MA}$	$7.14 * 10^{-2}$	mm^6/$cell^2$∗day	[44]

Table 4: Other coefficients used in the coupled model.

Parameter	Value	Unit	Reference
α_{MA}	10^{-3}	1/day	[44]
α_T	0.01	1/day	[44]
α_B	1.0	1/day	[44]
α_P	5.0	1/day	[44]
α_F	0.43	1/day	[44]
α_{MR}	4.0	1/day	Estimated
b_T	$1.7 * 10^{-2}$	mm^3/cell∗day	[44]
b_P	10^5	mm^6/$cell^2$∗day	[44]
b_P^B	$6.02 * 10^3$	mm^6/$cell^2$∗day	[44]
b_P^P	$2.3 * 10^6$	mm^3/cell∗day	[44]
ρ_T	2.0	—	[44]
ρ_B	16.0	Cell/mm^3	[44]
ρ_P	3.0	—	[44]
ρ_F	$5.1 * 10^4$	—	[44]
V_{LN}	160	Cells	Estimated

Our implementation is based on the finite difference method for the spatial discretization and the explicit method for the time evolution with an upwind scheme for the convective term of the equations. The upwind scheme discretizes the hyperbolic part of the PDEs using a bias for the flux direction given by the signal of the characteristic speeds. The upwind scheme uses an adaptive or solution-sensitive stencil to precisely simulate the direction of information propagation.

```
(1) begin
(2)    Initialize();
(3)    createFiles();
(4)    while ((t < iterPerDay ∗ numDays) && (A > tol)) do
(5)      calcIntegrals();
(6)      solveODEs(); // systemic response model
(7)      solvePDEs(); // local response model
(8)      recordResults();
(9)    end
(10) end
```

Algorithm 1: Main program for the coupled models.

Below there is an example of a finite difference operator used in the discretization of the Laplace operator that simulates the diffusion phenomenon in 3D:

$$
\begin{aligned}
&D_O\left(\frac{\partial^2 O(x,y,z)}{\partial x^2}+\frac{\partial^2 O(x,y,z)}{\partial y^2}+\frac{\partial^2 O(x,y,z)}{\partial z^2}\right)\\
&\approx D_O * \Big(\big(o[x+1,y,z]-2*o[x,y,z]+o[x-1,y,z]\big)\\
&\qquad \times\big(\Delta X^2\big)^{-1}\Big)\\
&+ D_O * \Big(\big(o[x,y+1,z]-2*o[x,y,z]+o[x,y-1,z]\big)\\
&\qquad \times\big(\Delta Y^2\big)^{-1}\Big)\\
&+ D_O * \Big(\big(o[x,y,z+1]-2*o[x,y,z]+o[x,y,z-1]\big)\\
&\qquad \times\big(\Delta Z^2\big)^{-1}\Big).
\end{aligned}
\tag{14}
$$

In (14), O represents the discretization of some types of cells, such as resting and activated macrophages; D_O is the diffusion coefficient of these populations of cells, x, y, and z are the position in the space, and ΔX, ΔY, and ΔZ are the space discretization.

The code was implemented using the C programming language and it was considered a 10×10 uniform grid with space discretization of $\Delta x = \Delta y = \Delta z = 0.1$ representing 10 mm^3 of tissue. The time step for both ODEs and PDEs is $\Delta t = 10^{-4}$ and the set of ODEs was solved with explicit Euler method. Algorithm 1 shows the general implementation of the coupling of models in which $iterPerDay = 10^4$, $numDays$ depends on the scenario simulated—for the coupled model it was considered equal to 30. A is the amount of antigens still present in the tissue. The simulation finishes when its value is less than or equal to a given threshold value, in this case $tol = 10^{-6}$.

An effort to parallelize the coupling of models is still in progress [43] and does not rely in the scope of this paper. More information about the implementation of the PDEs can be obtained in [40].

4. Results and Discussion

To show the importance of immune cells, molecules, and processes in the dynamics of the immune response and also to validate the coupling of models, a set of simulations were performed under distinct scenarios. The simulations start with a simple scenario were the cells of the IS are not considered (Case 1). Aiming to analyze the importance of the antibodies to the elimination of bacteria, the response is represented firstly by the simulation of only the innate cells (Case 2) and then with the complete model including activation of the lymphocytes, production, and migration of antibodies to the tissue (Case 3). A sensitivity analysis was performed in order to evaluate the behavior of the coupled mathematical model.

4.1. Case 1: Antigen Behavior. The purpose of this case is to show the diffusive term in the antigen Equation (1).

Initially, the antigen is injected only in the middle portion of the three-dimensional domain (Figure 4(a)). All images in Figure 4 show a cut view of the volume along the x-axis in order to better visualize both the initial condition and the diffusion of the antigens. The simulation shows that, without the immune system cells, after a few hours the antigen starts to spread over the domain because of the diffusion (Figure 4(b)). Furthermore, the replication can also be observed with the increase of the population of *S. aureus* (Figures 4(c) and 4(d)). Before the end of the 20-day simulation it can be observed that there are antigens all over the domain in considerable amounts. In this case only (1) was simulated without considering any kind of response. The cubic domain was sliced in half to improve the visualization.

Figures 5(a) and 5(b) show the logistical growth of antigen limited only by the space available, respectively, during 20 and 100 days. It can be observed that after 10 days the antigen reaches the maximum amount the simulated tissue could carry, which is assumed to be 50 antigens per cubic millimeter.

4.2. Case 2: Innate Response to an Antigen. Initially, without considering the trigger of the acquired response, the bacteria *S. aureus* increases for days held only by macrophages and the available space, which is assumed to contain at most 50 antigens per cubic millimeter. But after this initial period of time, macrophages are capable of restraining the antigen growth. However, they are not able to eliminate them completely. (Figure 6(a)).

Firstly, innate response to an antigen during 30 days was simulated (Figure 6(a)). According to the results shown by Figure 6(a), one can see that the amount of antigen increases until approximately 10 days (240 hours) and then decreases slowly. However, to understand what happens after those days that seemed to lead to an elimination of the antigen another test was performed for 100 days that showed that the initial amount of antigen injected in the tissue and the rates in which the innate immune cells arrive from the blood the antigen are not eliminated but stay on a chronic equilibrium state (Figure 6(b)). Throughout the simulations, the macrophages in the resting state come from the blood vessels which were positioned in the corners of the cubic domain along the y-axis (Figures 7(a) and 7(b)).

The set of Figures 8(a)–8(d) show the spread of the antigen over the tissue in a 10 mm^3 domain. Again, in order to better visualize the results the volume was cut along its x-axis. The initial amount of antigens was injected in the central portion of the tissue simulated (same used for Case 1) and it can be seen that the replication of the antigen and the absence of the acquired response lead to the spread of the antigen over the tissue restrained by the macrophages. The cubic domain was sliced in half to improve the visualization.

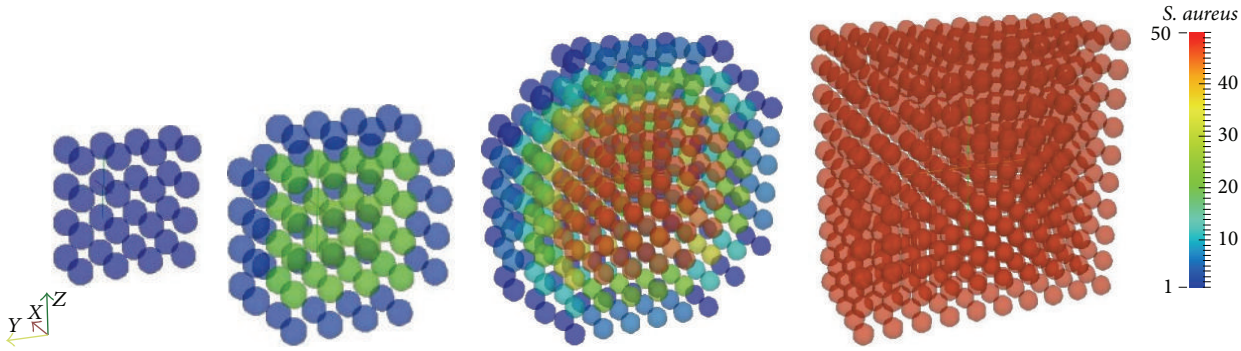
Figures 9(a) and 9(b) depict the averages of macrophages both resting and activated in the tissue, respectively, over 30 days of simulation. The state changing from resting to activated can be observed as the resting population decreases while the activated macrophage population increases, though they are not exactly opposite curves due to distinct phagocytosis and decay rates.

4.3. Case 3: Coupled Model. The third case represents the complete scenario with the APCs stimulating the lymphocytes. The coupled model scenario was simulated considering four blood capillaries and four lymph capillaries inside the 10 mm^3 cubic domain. The capillaries were placed, respectively, on the edges of the cube over the y-axis (Figure 7(b)) and distributed near the central portion of the cubic domain over the same y-axis (Figures 10(a) and 10(b)).

This distribution is given by a subroutine which is easily modified to allow distinct configurations. The antigen initial condition is the same simulated on the second case: an injection in the central portion of the domain. A 30-day simulation was performed and the results showing the coupling follow within this section. With the arrival of antibodies in the tissue the immune response was able to eliminate the antigen after 20 days of simulation as shown in Figure 11.

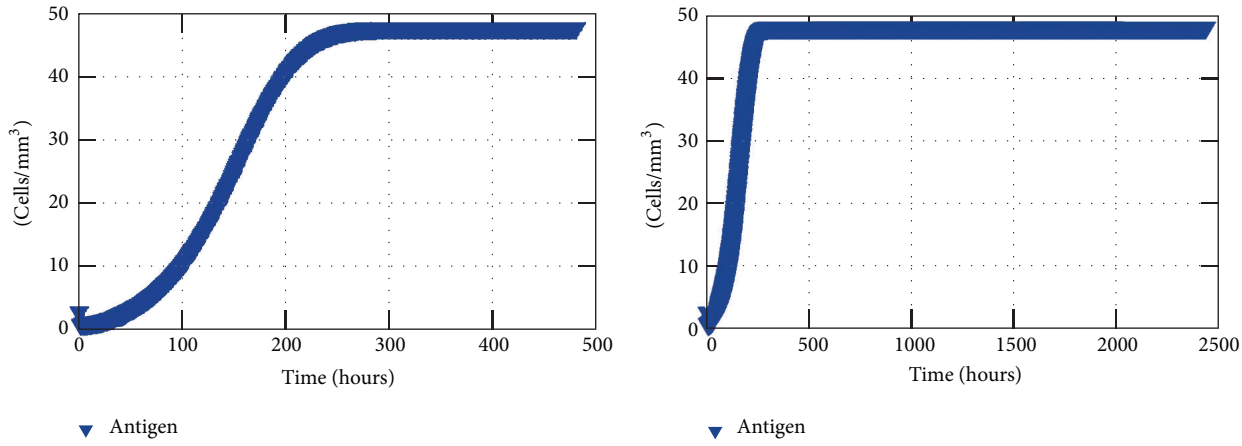
Figure 12 shows a comparison between the average population of activated macrophages inside the tissue and the population of activated macrophages in the LN. Logarithmic scale was used for the comparison. As one can observe, activated macrophages are migrating to the nearest LN to work as APCs: as the resting macrophages become activated, their population increases in the tissue as well as in the LN, which shows that the coupling is working.

Other feature of the coupling is the migration of antibodies produced after the stimulation cascade from the LN to the location where the infection takes place. The concentration of antibodies is shown in Figure 13 in which an average of antibodies inside the tissue is represented over time as well as the concentration of antibodies in the nearest LN. It can be observed that the antibodies start to be produced early in the simulation, just after a few hours, what is due to the presence of activated macrophages in the LN. A small amount of antibodies migrate to the local of the inflammation through the blood vessels after a few hours and a significant amount is present in the tissue after a few days which contributes to the elimination of antigens.



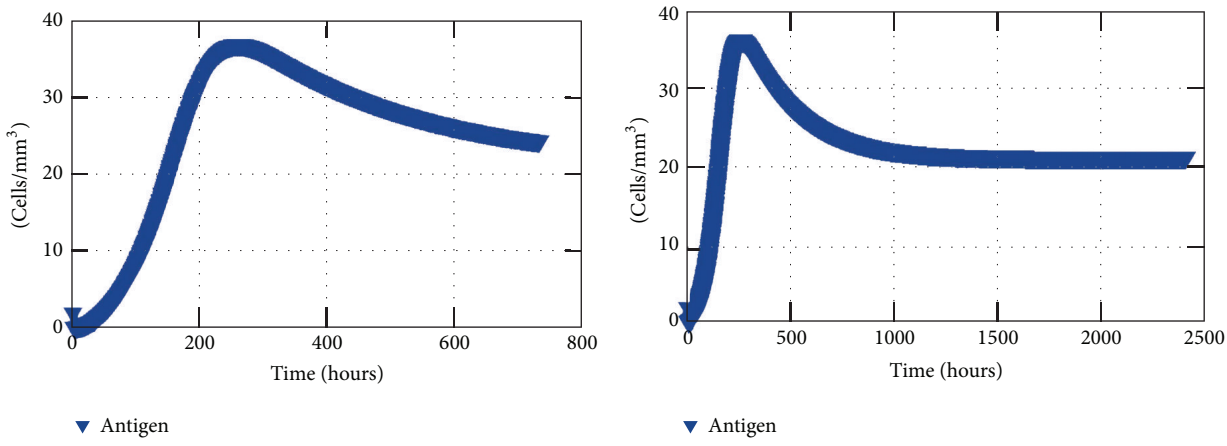
(a) Initial condition (b) After 48 hours (c) After 120 hours (d) After 320 hours

FIGURE 4: Initial condition and diffusion of the antigens at a 20-day simulation limited only by available space. After approximately 10 days of simulation the whole domain is filled with antigens.



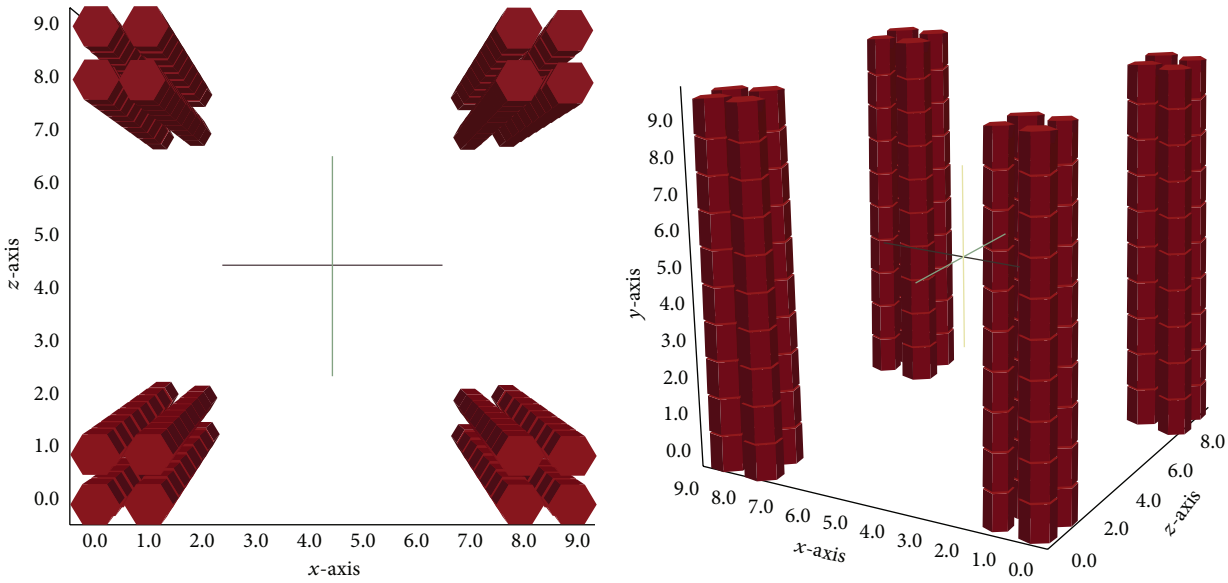
(a) 20 days (b) 100 days

FIGURE 5: Antigen average concentration inside the tissue without any response.



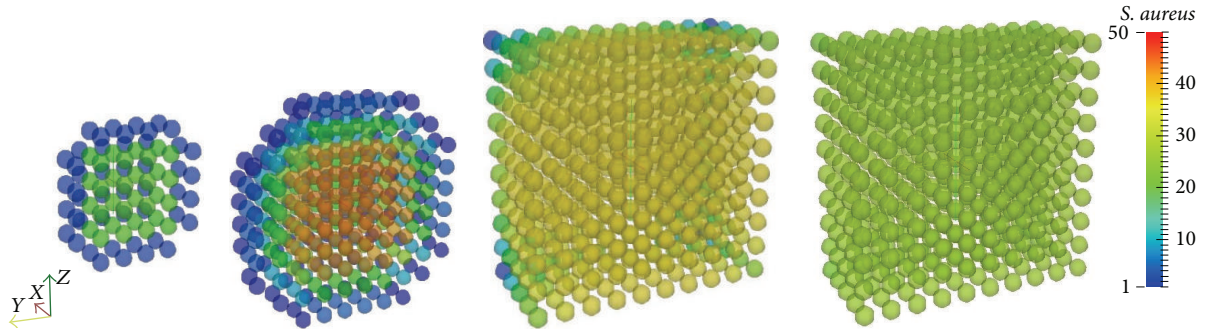
(a) 30 days of simulation (b) 100 days of simulation

FIGURE 6: The results show the average of antigen in each mm^3 of the tissue restrained only by the innate response.



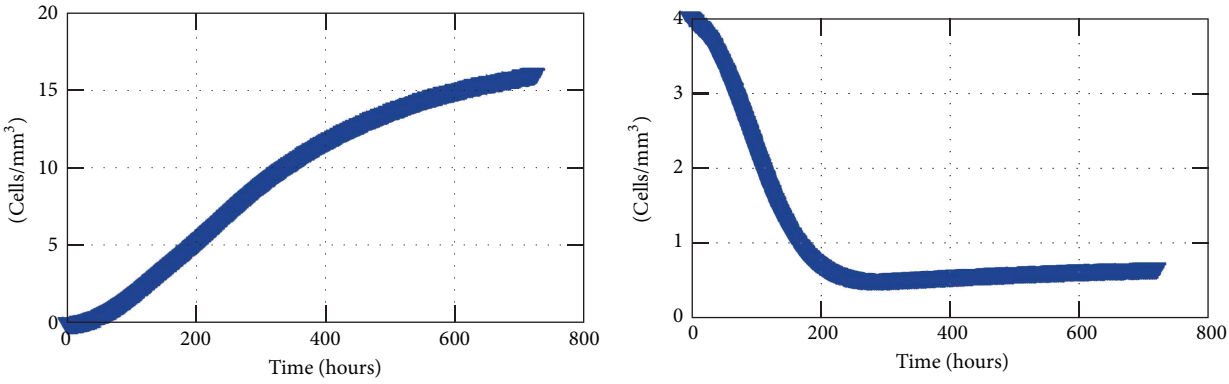
(a) X-Z plane distribution of the blood vessels (b) 3D distribution of the blood vessels

Figure 7: Scheme representing the position of the blood vessels on the edges of the tissue simulated.



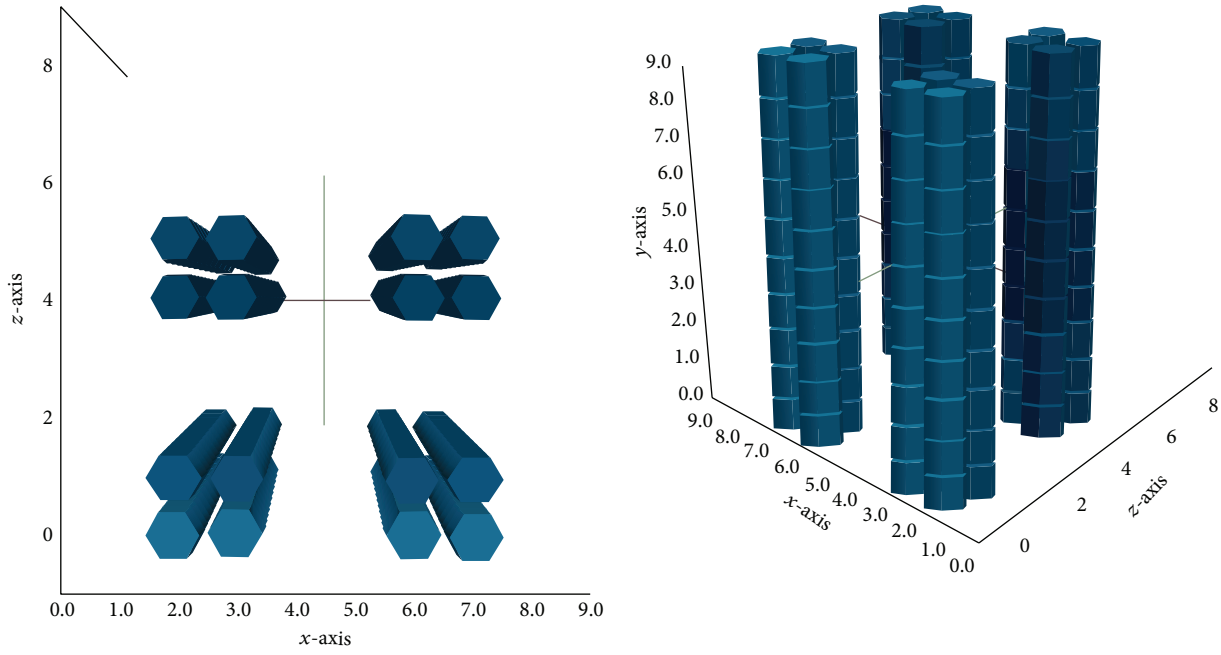
(a) After 48 hours (b) After 120 hours (c) After 240 hours (d) After 320 hours

Figure 8: Antigen diffusion during 30 days of simulation limited by the presence of macrophages. The macrophages are capable of restraining the antigen growth but they are not able to defeat this initial amount.



(a) Average of macrophages activated in the tissue (b) Average of macrophages resting in the tissue

Figure 9: Macrophages in the tissue over 30 days of simulation.



(a) X-Z plane distribution of the lymph vessels

(b) 3D distribution of the lymph vessels

FIGURE 10: Scheme representing the position of the lymph vessels in the tissue. They are not centralized to better visualization of the cells diffusion.

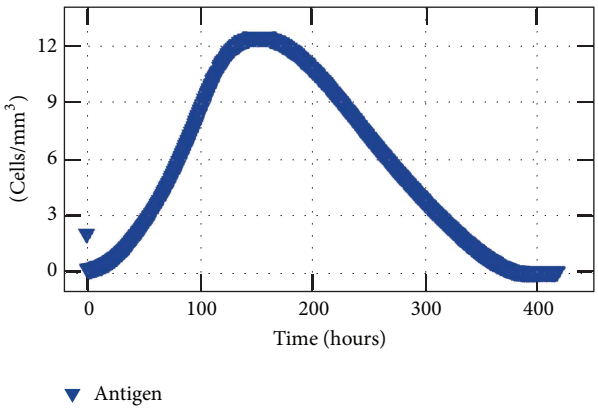


FIGURE 11: Antigen concentration (average) in the tissue for 20 days of simulation of the coupled model.

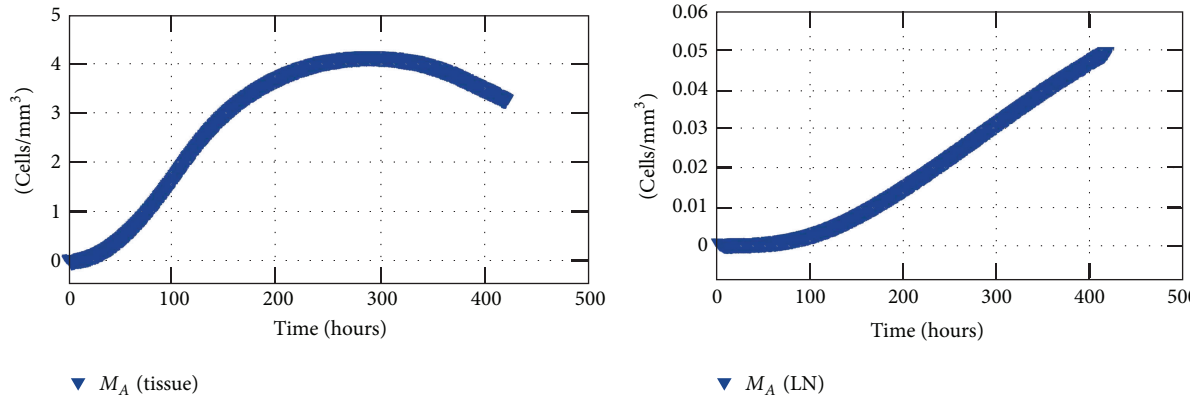
The average of antibodies, as well as the average of macrophages, tends to a steady state instead of decreasing after the elimination of the antigen. This happens due to the absence of a self-regulation process. At the moment, as the macrophages that are activated are still present in the LN, they continue to stimulate the lymphocytes to produce more antibodies which keep migrating to the tissue. However, the acquired response is self-regulated and we expect to add this feature to the model through the addition of anti-inflammatory cytokines, such as IL-4 or IL-10. This work is already in progress.

Figure 14 shows the concentration of T-Lymphocytes, B-Lymphocytes, and plasma cells, within 20 days of simulation. It could be noticed that the activation is happening early with the increase of lymphocytes in the first hours. Thus, the peak of those cells occurs approximately after 2 days for the T-lymphocytes and around the 5th day for B-lymphocytes and plasma cells.

Aiming to show the diffusive process and the effect of the arrival of the antibodies in the tissue the set of Figures 15(a)–15(f) presents the antibodies and Figures 16(a)–16(f) present the antigens. The antibodies arrive at the tissue through the lymph vessels which are positioned according to Figures 10(a) and 10(b). The domain is a hexahedron which was sliced to better visualization henceforth there are 4 vessels but Figures 15(a)–15(f) only show half of them. Moreover, Figures 16(a)–16(f) are also sliced to better visualization of the diffusion.

After a couple of days of the beginning of the simulation, it is possible to see the antigen starting to diffuse (Figure 16(a)). The antigen continues to diffuse slowly restrained by the innate response until approximately the 3rd day. After a few days the amount of antibodies that is arriving at the tissue helps macrophages to defeat the antigen more efficiently in the regions where there is more concentration of antibodies (Figures 16(c)–16(f)).

4.4. Sensitivity Analysis of the Coupled Model. The sensitivity analysis can be used to help with the verification of a mathematical model by evaluating how the model responds to changes in one or more inputs. The validation of the model involves comparison of the results to independent observations from the system being modeled which is not always feasible. Therefore, the sensitivity analysis can be used to understand the behavior of the model and reach a



(a) Activated macrophages in the tissue

(b) Activated macrophages in the LN

Figure 12: Average concentration of activated macrophages in the tissue and concentration of activated macrophages in the nearest LN.

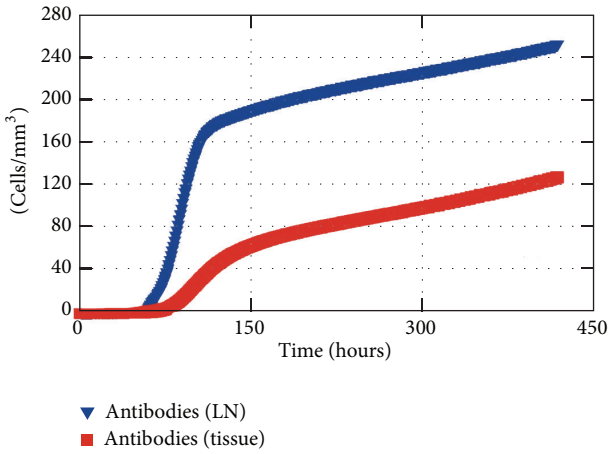


Figure 13: Antibodies in the LN and in the tissue (average) during 20 days of simulation.

comfortable position in terms of qualitative results [66]. Thus, there are several ways of performing this assessment of the sensitivity of the model; it was chosen herein the so called one-factor-at-a-time approach (OAT), which is the most used strategy [67].

The sensitivity analysis of the coupled model considered the complete scenario during 30 days in a cubic centimeter domain. Each chosen parameter on Table 5 was assessed one at a time varying its value from −100% to +200% to understand its influence on the output. Table 5 presents the chosen parameters with a brief description and their maximum error value (Max_{err}).

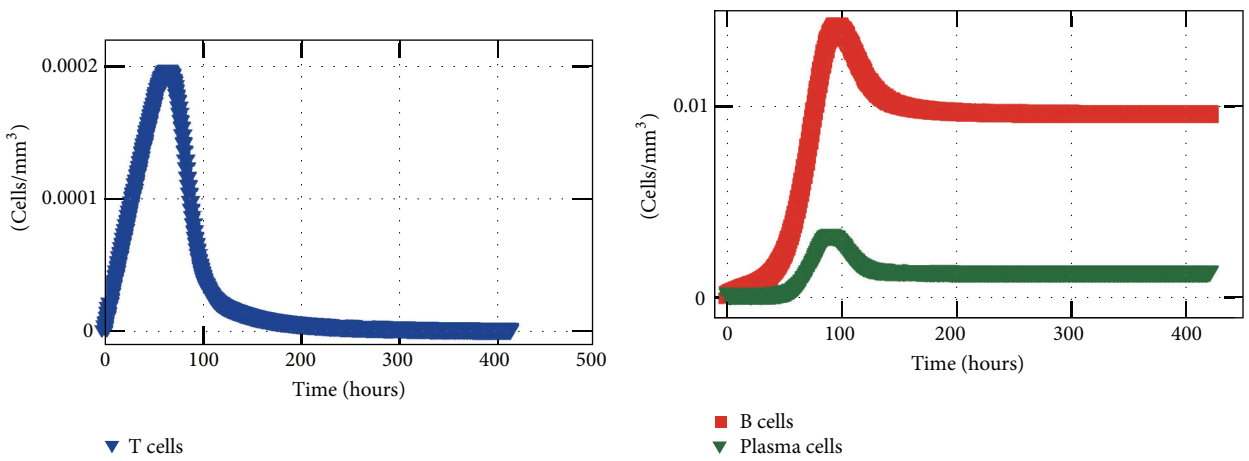
The error was calculated applying (15) to each parameter as follows:

$$\text{Max}_{\text{err}} = \text{MAX}_k \left(\frac{\sqrt{\sum_{i=0}^{N} \left(E_{\text{orig}}(i) - E_k(i) \right)^2}}{\sqrt{\sum_{i=0}^{N} \left(E_{\text{orig}} \right)^2}} \right), \quad (15)$$

Table 5: Sensitivity analysis—chosen parameters, description, and the maximum error value.

Parameter	Description	Max_{err}
γ_{AM}	Macrophage activation rate	6.47
α_F	Antibodies migration rate	6.36
M_{R_0}	Resting macrophage initial condition	6.17
β_A	Antigen replication rate	5.67
ρ_F	Antibodies release rate	5.15
k_A	Antigen carrying capacity coefficient	4.82
b_P^P	B cell expenditure to become plasma cell	4.09
α_{MA}	Activated macrophage migration rate	4.09
$\lambda_{AF\|MA}$	Activated macrophage phagocytosis rate of opsonized antigen	4.06
b_P	T cell expenditure to stimulate B cell	2.42
D_{MR}	Resting macrophage diffusion coefficient	0.93
D_A	Antigen diffusion coefficient	0.85
λ_{MA}	Activated macrophage phagocytosis rate	0.76
b_P^B	B cell stimuli coefficient	0.73
D_{MA}	Activated macrophage diffusion coefficient	0.24
A_0	Antigen initial condition	0.09
α_{MR}	Resting macrophage source coefficient	0.07
λ_{MR}	Resting macrophage phagocytosis rate	0.05
$\lambda_{AF\|MR}$	Resting macrophage phagocytosis rate of opsonized antigen	0.01
b_T	T cell stimuli coefficient	$3 * 10^{-4}$

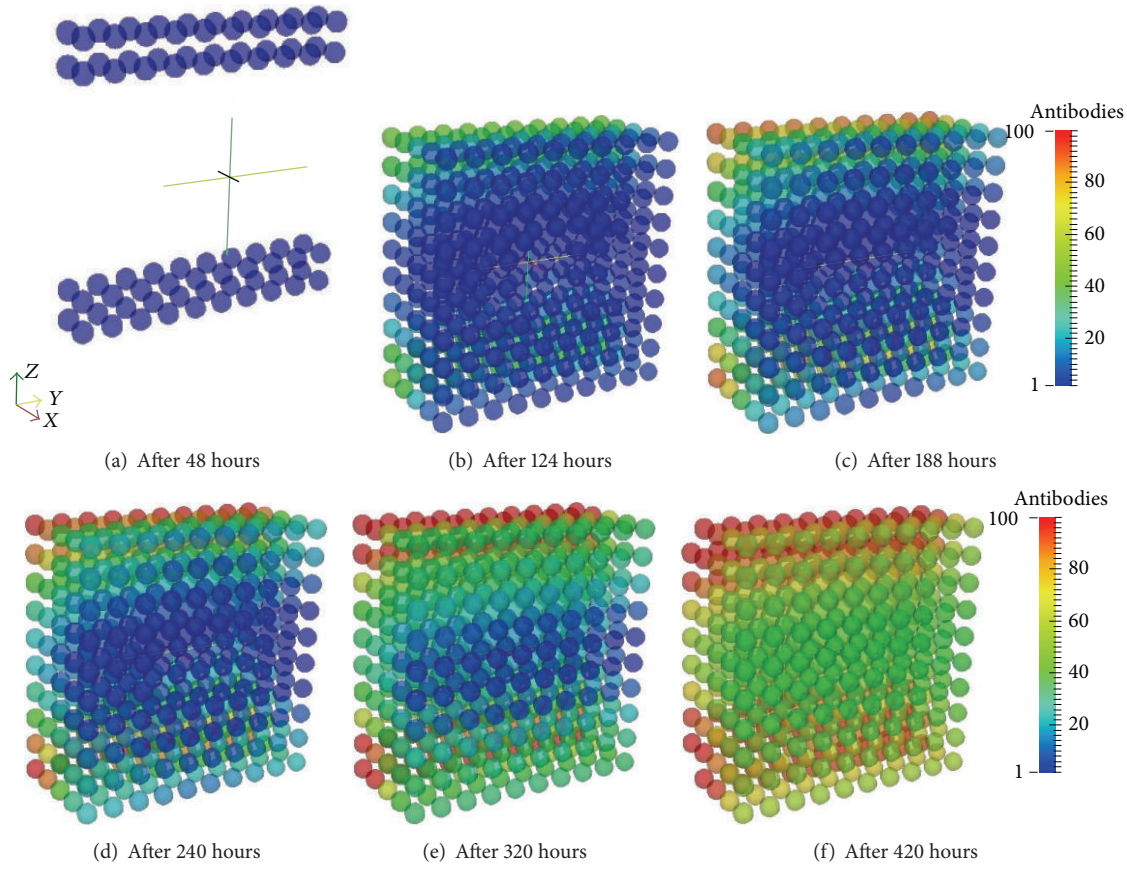
in which k index each variation within the same parameter. E_{orig} is the number of antigens over time using the original set of parameters, E_k represents each resultant number of antigens over time with the variation of one parameter at-a-time and N is the number of time steps. Thus, we have a maximum error value for each parameter that enables us to understand which parameters are the most sensitive of the coupled model.



(a) T cells in the LN

(b) B and plasma cells in the LN

FIGURE 14: T-lymphocyte, B-lymphocyte, and plasma cell concentrations in the LN.



(a) After 48 hours

(b) After 124 hours

(c) After 188 hours

(d) After 240 hours

(e) After 320 hours

(f) After 420 hours

FIGURE 15: Antibodies arriving in the tissue through the blood vessels. Initially the antibodies do not exist in the tissue and they arrive through the capillaries and start to diffuse over the tissue.

Among the 20 assessed parameters we are going to consider those with $\text{Max}_{\text{err}} \geq 1$. Therefore, the most sensitive parameter is the rate of activation of macrophages γ_{AM}. Without the activation, the acquired response is not triggered and the response depends only on the resting macrophages (Figure 17(a)). The second one is the migration rate of the antibodies to the tissue (α_F). This parameter is essential to the acquired response and without this migration rate there are only resting and activated macrophages in the tissue (Figure 17(b)). The initial condition of resting macrophages (M_{R_0}) is also significant for the model by the fact that without them there is practically no response as they are responsible

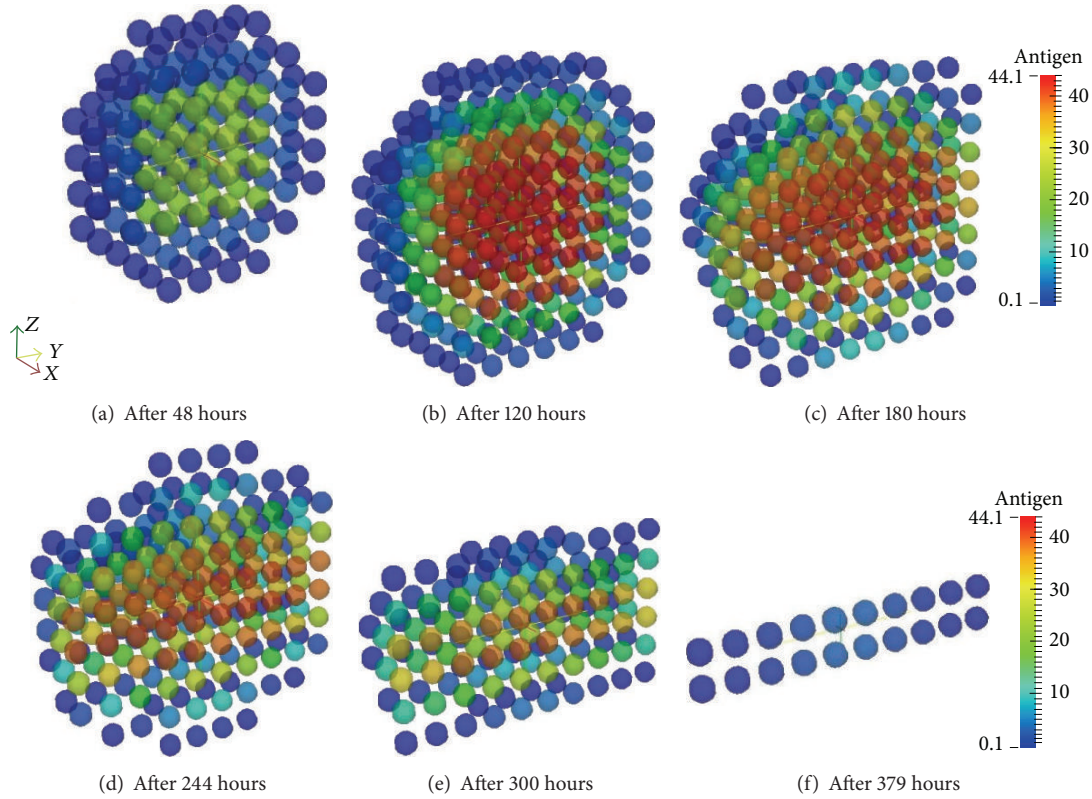
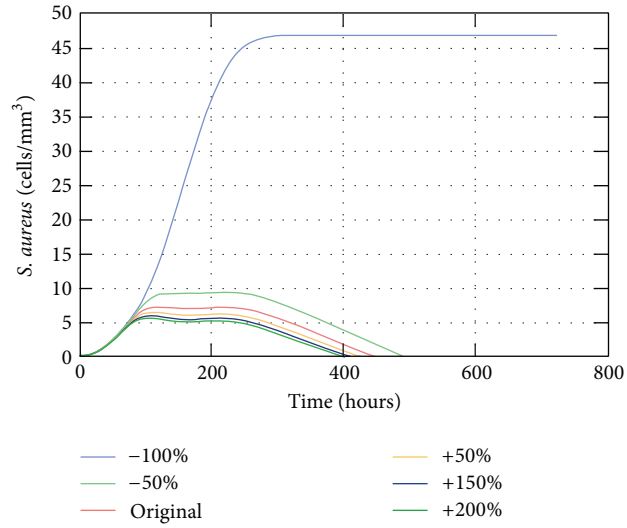


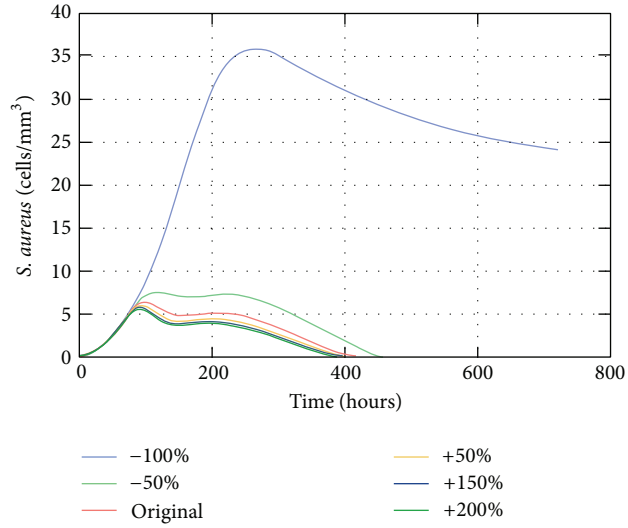
FIGURE 16: Antigen diffusion during 20 days of simulation of the coupled model. The macrophages are capable of restraining antigen growth but they are not able to defeat this initial amount. However, the antibodies arriving through the blood vessels help macrophages to eliminate the antigen through opsonization.

for recognizing and engulfing the antigen (Figure 17(c)). After they identify the antigens they become APCs which triggers the acquired response. According to the results shown in Figure 17(d), without the replication rate (−100%) the amount of antigen in the tissue remains reduced but sufficient to trigger the immune response which eliminates the antigen. Increasing this rate, the immune response as a whole takes more time to eliminate the antigen. With more than 50% of increase the response is not able to defeat the antigen in 30 days. The release rate of antibodies in the LN (ρ_F) also affects directly the acquired response. As it can be observed in Figure 17(e), doubling the original value the antigens are eliminated in almost half the time (approximately 9 days) whereas decreasing this rate by 50% the response takes approximately 28 days. The carrying capacity parameter, shown in Figure 17(f), is the only one that cannot be varied from −100%; otherwise it would generate a division by zero (1). So, it was varied from −50% and for that value, the response is able to eliminate the antigen a couple of days before the original value used in simulations, whereas doubling the value, it takes approximately more 10 days to defeat the antigen. If there is more antigen that is able to replicate the immune response needs more time to defeat them.

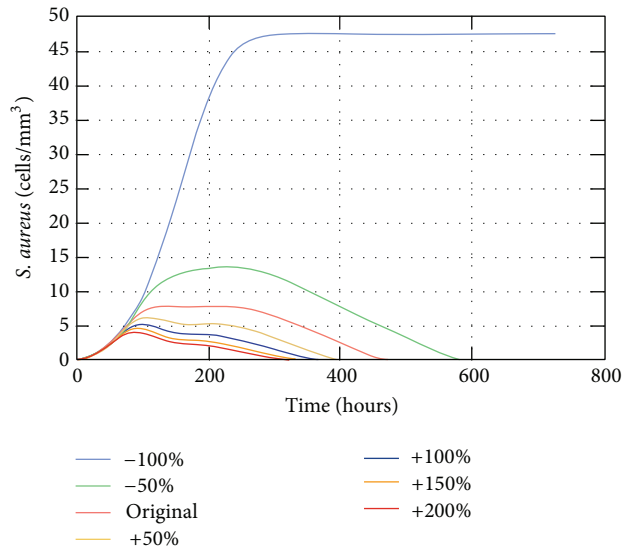
The coefficient b_P^P is important to determine the number of B cells that turn into plasma cells. If this value is reduced to 100% it means that not a single B cell turns into a plasma cell leading to a nonexistent specific response. However, increasing this parameter leads to larger production of antibodies and quicker specific response to eliminate antigens (Figure 18(a)). Moreover, the activated macrophage migration rate to the LN (α_{MA}) is essential to initiate the acquired response as without this migration the antigen is not presented to the lymphocytes in the LN and the specific antibodies are not mass produced. Thus, if this parameter is set to zero there is only the innate response (Figure 18(b)). The activated macrophages phagocytosis rate of opsonized antigen, $\lambda_{AF|MA}$, also shows great discrepancy if set to zero (Figure 18(c)). This rate is important to the effectiveness of the acquired response and the more its value is increased the earlier the response is able to eliminate the antigens. Setting the coefficient b_p equal to zero means that the B cells are stimulated even without the presence of T cells, leading to the constant stimulation of existent B cells in the LN. Those B cells turn into plasma cells which population increases promptly in the LN. Also, as soon as the activated macrophages arrive in the LN, a large production of antibodies starts. As a consequence, a great number of antibodies arrive in the tissue, approximately 4 days after the injection of antigens.



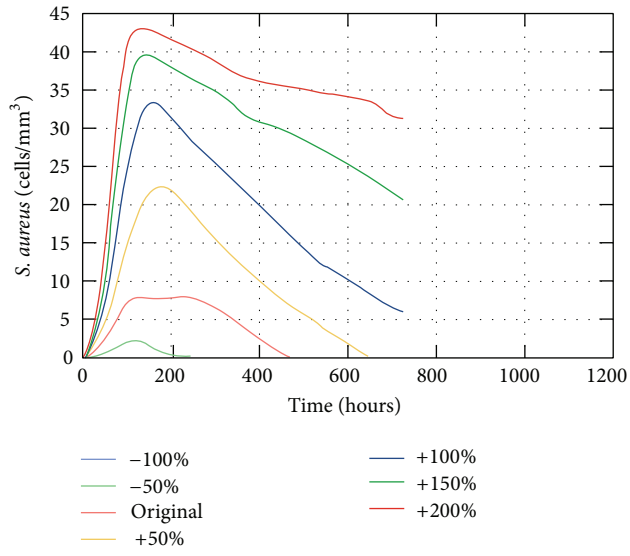
(a) Macrophage activation (γ_{AM})



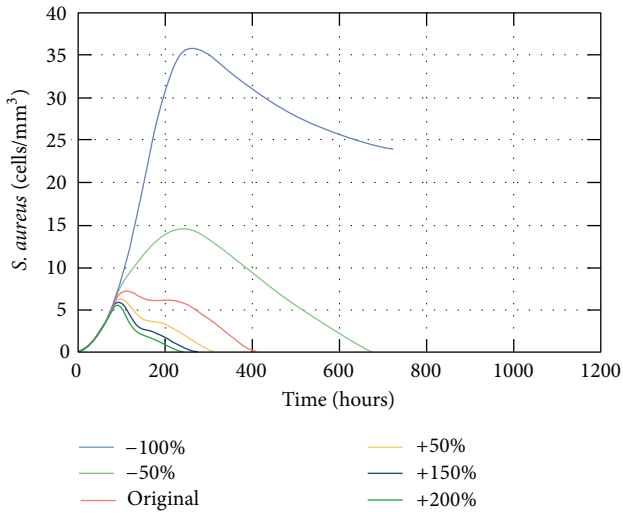
(b) Antibodies migration rate (α_F)



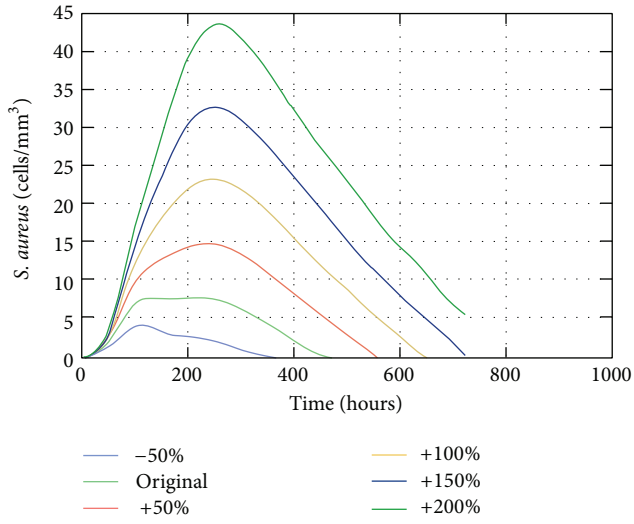
(c) Resting macrophage initial condition (M_{R_0})



(d) Antigen replication rate (β_A)



(e) Antibodies release rate (ρ_F)



(f) Antigen carrying capacity (k_A)

FIGURE 17: Impact of the variation of γ_{AM}, α_F, MR_0, β_A, ρ_F and k_A on the population of antigens.

Approximately 3 days after the arrival of the antibodies the antigens are eliminated. Meanwhile, if the value of that coefficient is 200% bigger, it means that a lot more T cells are needed to stimulate the B cells, leading to a much smaller number of antibodies arriving in the tissue, which in turn are not able to help macrophages to defeat the antigens in 30 days (Figure 18(d)).

4.5. Discussion. Based upon two distinct mathematical models of the human immune system a novel form of coupling models was developed. This coupling was performed firstly by analyzing the features of each previous model to identify the possible bridges between them and then building the linking itself. The choice of coupling models instead of developing a whole new set of equations is due to the fact that these models were already validated experimentally, which is not easy to achieve without collaborative work. We chose a model with already fitted parameters [44] and we had to convert those parameters to the unit we were using in previous work [42] as the former units were given in molar concentration (mol) and we expect to analyse the number of interactive cells per cubic millimeter.

In order to perform the coupling we had to add some terms to the PDEs and even a new Equation (4) to represent the interactions between the cells and their migration. The former PDEs based model [42] was not able to simulate the presence of antibodies inside the tissue which was achieved with the coupling performed in this work. That model represents other features of the immune response as the presence of neutrophils and the chemotaxis process which are not present in this coupling yet due to simplification. Our intention is to focus on the coupling and then improve the model with those characteristics of the innate IS. The terms that were added to the model are the ones which control the flux of cells between the tissue and the nearest LN (3) and the terms that represent the opsonization process in (1).

The former ODEs based model [44] actually represented some features of the acquired response such as the clonal expansion of T helper cells with delay differential equations (DDEs). DDEs have been used to model biological processes as they give a better approximation for such aspects [68, 69]. We have opted not to use DDEs initially in order to simplify the model as we solve the equations using our own solver. We hope to introduce this concept in the future to better represent those biological processes. We modified that model in the following aspects: (a) we removed the equation for bacteria equilibrium due to the fact that we are representing this insertion of bacteria inside the tissue (modeled by the PDE given by (1)) and (b) we modified the equation that represents the antigen presenting cells in the LN (5). The previous equation considered bacteria stimulation and natural decay. The modified equation solely considers the flux of active cells between the tissue and the LN as the activation and natural decays are represented locally in the tissue (3). This representation required an integration of the concentration of cells in the tissue to estimate the amount of cells in the LN (7).

We would like to reinforce that those two models presented herein were chosen due to the availability of parameters already fitted. The outcomes of the coupled model agree qualitatively with the literature [44]. Further quantitative validation of the coupling is still required. That could be achieved, for example, by comparing the outputs of the coupled model to results obtained experimentally. This is an ongoing work. We would also like to perform this coupling with other models to gain insights into a specific infection scenario.

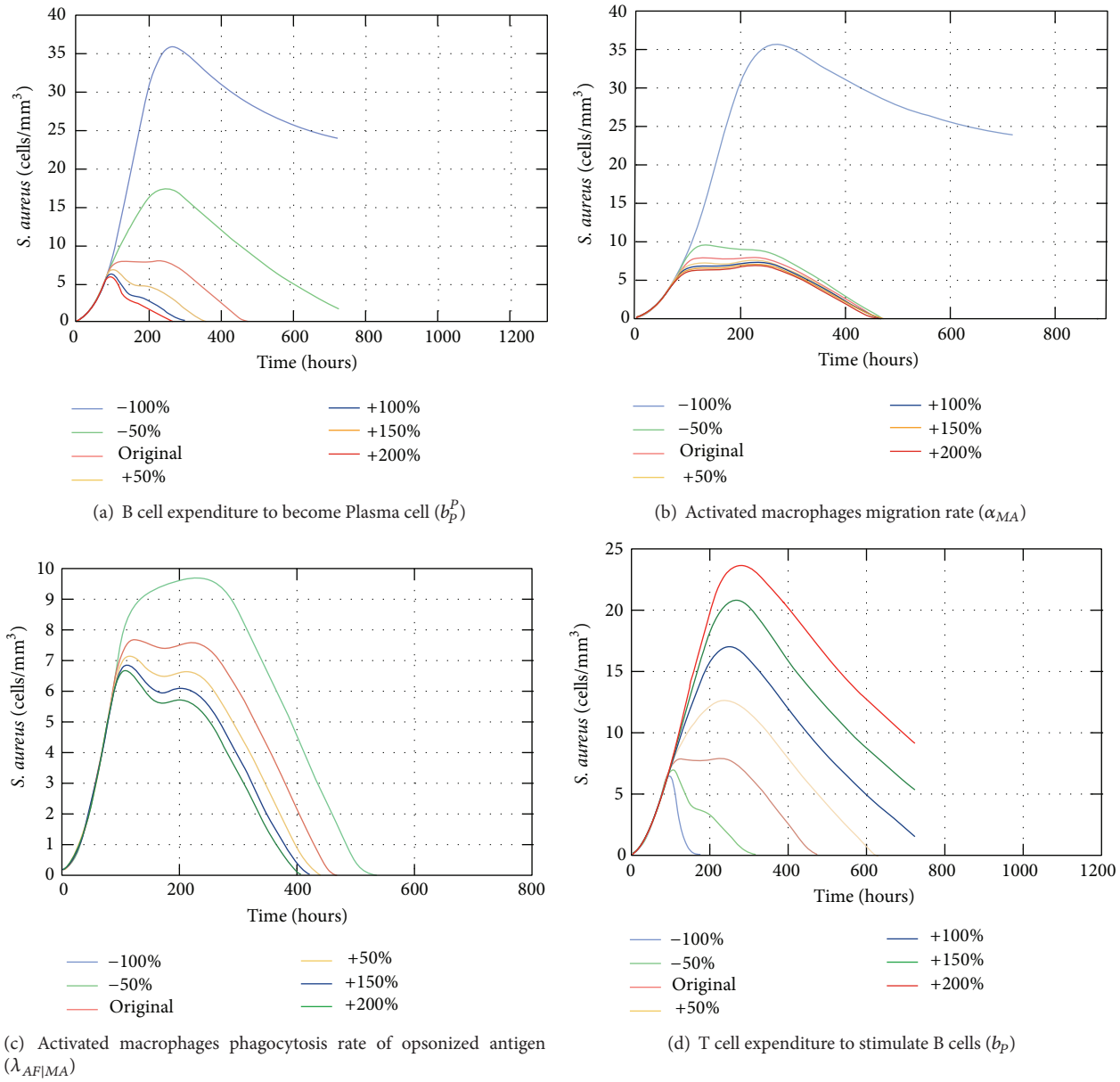
5. Conclusions

This work presented the coupling of two distinct models of different aspects of the immune system: one of them uses PDEs to model the dynamics of cells in a three-dimensional section of tissue and the other one uses ODEs to model the dynamics of cells in the nearest LN. To the best of our knowledge, the integration of two models in the format presented has not been proposed before in the field of immune systems. To exemplify the coupling, a mathematical and computational coupling of models was presented that simulates the immune response to *S. aureus* bacteria into a three-dimensional section of a tissue. To achieve this goal, the models reproduce the initiation, maintenance, and resolution of innate and adaptive immune response. A set of PDEs and ODEs are used to model the main agents involved in this processes, like the antigen, macrophages, antibodies, and T and B cells.

The model presented in this work represents an infection scenario: the diffusion of antigens into the tissue and the migration of macrophages to combat the infection. Macrophages also migrate outside the tissue and stimulate the adaptive IS to produce antibodies, which in turn migrate inside the tissue and opsonize the antigens. The proposed integrated model was capable of reproducing qualitatively the spatial and temporal behavior of resting and activated macrophages as well as specific antibodies.

A sensitivity analysis was performed for the coupled model showing that the most relevant parameters are the ones related to the activation of the response, as the macrophage activation rate (γ_{AM}), effectiveness of the acquired response as the antibodies migration rate (α_F), and the presence of the immune responses itself as the resting macrophage initial condition (MR_0). Other parameters that are important to the success of the immune response are the antigen replication rate (β_A), antibodies release rate (ρ_F), and the amount of antigens that could grow in the tissue (k_A).

We expect that with that spatial coupled model we could simulate and analyze the evolution of damages caused to an organ parenchyma, for example, the damage in the lung tissue caused by tuberculosis or pneumonia. Also, we are already implementing a more complete mathematical model including molecules, like cytokines, and others processes



(a) B cell expenditure to become Plasma cell (b_P^P)

(b) Activated macrophages migration rate (α_{MA})

(c) Activated macrophages phagocytosis rate of opsonized antigen ($\lambda_{AF|MA}$)

(d) T cell expenditure to stimulate B cells (b_P)

FIGURE 18: Impact of the variation of b_P^P, α_{MA}, $\lambda_{AF|MA}$, and b_P on the population of antigens.

involved in the immune responses to consider the chemotaxis process. Furthermore, we intend to improve the visualization of the damage caused to the tissue in order to compare to medical imaging results. Thus, we believe that the coupling of models in the proposed format could provide some insight into the behavior of the immune system.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to express their thanks to CAPES, CNPq, FAPEMIG, and UFJF (Programa de Apoio à Publicação/Pró-reitoria de Pesquisa/Universidade Federal de Juiz de Fora (PROPESQ/UFJF)) for funding this work.

References

[1] H. Kitano, "Systems biology: a brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[2] L. Sompayrac, *How the Immune System Works*, Blackwell, 3rd edition, 2008.

[3] C. Shi and E. G. Pamer, "Monocyte recruitment during infection and inflammation," *Nature Reviews Immunology*, vol. 11, no. 11, pp. 762–774, 2011.

[4] D. F. Singer and J. J. Linderman, "The relationship between antigen concentration, antigen internalization, and antigenic complexes: Modeling insights into antigen processing and presentation," *The Journal of Cell Biology*, vol. 111, no. 1, pp. 55–68, 1990.

[5] N. G. B. Agrawal and J. J. Linderman, "Calcium response of helper T lymphocytes to antigen-presenting cells in a single-cell assay," *Biophysical Journal*, vol. 69, no. 3, pp. 1178–1190, 1995.

[6] J. Carneiro, A. Coutinho, J. Faro, and J. Stewart, "A model of the immune network with B-T cell co-operation. I—Prototypical structures and dynamics," *Journal of Theoretical Biology*, vol. 182, no. 4, pp. 513–529, 1996.

[7] A. S. Perelson and P. W. Nelson, "Mathematical analysis of HIV-1 dynamics in vivo," *SIAM Review*, vol. 41, no. 1, pp. 3–44, 1999.

[8] C. Keşmir and R. J. de Boer, "A mathematical model on germinal center kinetics and termination," *Journal of Immunology*, vol. 163, no. 5, pp. 2463–2469, 1999.

[9] M. L. Gatton and Q. Cheng, "Modeling the development of acquired clinical immunity to *Plasmodium falciparum* malaria," *Infection and Immunity*, vol. 72, no. 11, pp. 6538–6545, 2004.

[10] C. E. Warrender, *Modeling intercellular interactions in the peripheral immune system [Ph.D. thesis]*, The University of New Mexico, Albuquerque, NM, USA, 2004.

[11] D. L. Chao, M. P. Davenport, S. Forrest, and A. S. Perelson, "A stochastic model of cytotoxic T cell responses," *Journal of Theoretical Biology*, vol. 228, no. 2, pp. 227–240, 2004.

[12] C. Lundegaard, O. Lund, C. Keşmir, S. Brunak, and M. Nielsen, "Modeling the adaptive immune system: predictions and simulations," *Bioinformatics*, vol. 23, no. 24, pp. 3265–3275, 2007.

[13] A. M. Smith and R. M. Ribeiro, "Modeling the viral dynamics of influenza a virus infection," *Critical Reviews in Immunology*, vol. 30, no. 3, pp. 291–298, 2010.

[14] A. S. Perelson and G. Weisbuch, "Immunology for physicists," *Reviews of Modern Physics*, vol. 69, no. 4, pp. 1219–1267, 1997.

[15] V. Narang, J. Decraene, S.-Y. Wong et al., "Systems immunology: a survey of modeling formalisms, applications and simulation tools," *Immunologic Research*, vol. 53, no. 1–3, pp. 251–265, 2012.

[16] W. Wakeland, L. Macovsky, and G. An, "A hybrid simulation model for studying acute inflammatory response," in *Proceedings of the Spring Simulation Multiconference (SpringSim '07)*, vol. 2, pp. 39–46, Society for Computer Simulation International, San Diego, Calif, USA, 2007.

[17] I. O. Knop, A. B. Pigozzo, B. M. Quintela et al., "Modeling human immune system using a system dynamics approach," in *5th European Conference of the International Federation for Medical and Biological Engineering*, A. Jobbagy, R. Magjarevic, and R. Magjarevic, Eds., vol. 37 of *IFMBE Proceedings*, pp. 363–366, Springer, Berlin, Germany, 2012.

[18] F. Celada and P. E. Seiden, "A computer model of cellular interactions in the immune system," *Immunology Today*, vol. 13, no. 2, pp. 56–62, 1992.

[19] M. Bezzi, *Modeling Evolution and Immune System by Cellular Automata*, Rivista del Nuovo Cimento, Editrice Compositori, 2001.

[20] R. M. Zorzenon Dos Santos and S. Coutinho, "Dynamics of HIV infection: a cellular automata approach," *Physical Review Letters*, vol. 87, no. 16, Article ID 168102, 2001.

[21] P. M. A. Sloot, F. Chen, and C. Boucher, "Cellular automata model of drug therapy for hiv infection," in *Proceedings of the 5th International Conference on Cellular Automata for Research and Industry (ACRI '02)*, pp. 282–293, Springer, London, UK, 2002.

[22] R. Puzone, B. Kohler, P. Seiden, and F. Celada, "IMMSIM, a flexible model for in machina experiments on immune system responses," *Future Generation Computer Systems*, vol. 18, no. 7, pp. 961–972, 2002.

[23] F. Castiglione and M. Bernaschi, "C-immsim: playing with the immune response," in *Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS '04)*, Leuven, Belgium, July 2004.

[24] M.-S. Martin and G. Mack, "Simmune, a tool for simulating and analyzing immune system behavior," CoRR, 2008.

[25] N. R. Jennings, "Agent-based computing: promise and perils," in *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI '99)*, vol. 2, pp. 1429–1436, Stockholm, Sweden, August 1999.

[26] A. Grilo, A. Caetano, and A. Rosa, "Agent-based artificial immune system," in *Proceedings of the Genetic and Evolutionary Computation Conference Late Breaking Papers*, San Francisco, Calif, USA, 2001.

[27] G. An and I. A. Lee, "Computer simulation to study inflammatory response," *Simulation and Gaming*, vol. 32, no. 3, pp. 344–361, 2001.

[28] Y. Vodovotz, C. Chow, J. Bartels et al., "Mathematical simulation of sepsis and trauma," in *Proceedings of the 11th Congress of the European Shock Society (ESS '05)*, Vienna, Austria, January 2005.

[29] S. Bandini, S. Manzoni, and G. Vizzari, "Immune system modelling with situated cellular agents," in *Proceedings of the 1st International Workshop on Multi-Agent Systems for Medicine, Computational Biology, and Bioinformatics and 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '05)*, Utrecht , The Netherlands, July 2005.

[30] V. Baldazzi, F. Castiglione, and M. Bernaschi, "An enhanced agent based model of the immune system response," *Cellular Immunology*, vol. 244, no. 2, pp. 77–79, 2006, International Conference on Immunogenomics and Immunomics , Budapest, Hungary, October 2006.

[31] Z. Guo, P. M. A. Sloot, and J. C. Tay, "A hybrid agent-based approach for modeling microbiological systems," *Journal of Theoretical Biology*, vol. 255, no. 2, pp. 163–175, 2008.

[32] X. Dong, P. T. Foteinou, S. E. Calvano, S. F. Lowry, and I. P. Androulakis, "Agent-based modeling of endotoxin-induced acute inflammatory response in human blood leukocytes," *PLoS ONE*, vol. 5, no. 2, Article ID e9249, 2010.

[33] D. Kirschner, *The Multi-Scale Immune Response to Pathogens: M. Tuberculosis as an Example*, Springer, 2010.

[34] M. A. Possi, A. P. Oliveira, C. M. G. Chaves, F. R. Cerqueira, and J. E. C. Arroyo, "An insilico immune system model for investigating human autoimmune diseases," in *37th Conferencia Latinoamericana de Informatica (CLEI '11)*, Quito, Equador, 2011.

[35] V. A. Folcik, G. Broderick, S. Mohan et al., "Using an agent-based model to analyze the dynamic communication network of the immune response," *Theoretical Biology and Medical Modelling*, vol. 8, no. 1, article 1, 2011.

[36] M. Niazi and A. Hussain, "Agent-based computing from multi-agent systems to agent-based models: a visual survey," *Scientometrics*, vol. 89, no. 2, pp. 479–499, 2011.

[37] J. C. Tay and A. Jhavar, "Cafiss: a complex adaptive framework for immune system simulation," in *Proceedings of the 20th Annual ACM Symposium on Applied Computing*, pp. 158–164, ACM Press, New York, NY, USA, March 2005.

[38] K. Reed, K. Schalla, J. Tran et al., "General model of the innate immune response," Tech. Rep. 2011-03, 2011.

[39] B. Su, W. Zhou, K. S. Dorman, and D. E. Jones, "Mathematical modelling of immune response in tissues," *Computational and Mathematical Methods in Medicine*, vol. 10, no. 1, pp. 9–38, 2009.

[40] A. B. Pigozzo, G. C. Macedo, R. W. dos Santos, and M. Lobosco, "Implementation of a computational model of the innate immune system," in *Artificial Immune Systems*, vol. 6825 of *Lecture Notes in Computer Science*, pp. 95–107, 2011.

[41] D. E. Kirschner, S. T. Chang, T. W. Riggs, N. Perry, and J. J. Linderman, "Toward a multiscale model of antigen presentation in immunity," *Immunological Reviews*, vol. 216, no. 1, pp. 93–118, 2007.

[42] A. B. Pigozzo, G. C. Macedo, R. W. dos Santos, and M. Lobosco, "On the computational modeling of the innate immune system," *BMC Bioinformatics*, vol. 14, supplement 6, article S7, 2013.

[43] P. A. F. Rocha, M. P. Xavier, A. B. Pigozzo et al., "A three-dimensional computational model of the innate immune system," in *Computational Science and Its Applications—ICCSA 2012*, B. Murgante, O. Gervasi, S. Misra et al., Eds., vol. 7333 of *Lecture Notes in Computer Science*, pp. 691–706, Springer, Berlin, Germany, 2012.

[44] G. I. Marchuk, *Mathematical Modelling of Immune Response in Infectious Diseases*, vol. 395 of *Mathematics and Its Applications*, Kluwer Academic Publishers, 1997.

[45] S. Marino, A. Myers, and J. L. a. Flynn, "TNF and IL10 are major factors in modulation of the phagocytic cell environment in lung and lymph node in tuberculosis: a next-generation two-compartmental model," *Journal of Theoretical Biology*, vol. 265, no. 4, pp. 586–598, 2010.

[46] A. M. Smith, J. A. McCullers, and F. R. Adler, "Mathematical model of a three-stage innate immune response to a pneumococcal lung infection," *Journal of Theoretical Biology*, vol. 276, pp. 106–116, 2011.

[47] G. Nieman, D. Brown, J. Sarkar et al., "A two-compartment mathematical model of endotoxin-induced inflammatory and physiologic alterations in swine," *Critical Care Medicine*, vol. 40, no. 4, pp. 1052–1063, 2012.

[48] J. Day, A. Friedman, and L. S. Schlesinger, "Modeling the immune rheostat of macrophages in the lung in response to infection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 27, pp. 11246–11251, 2009.

[49] R. Kumar, C. C. Chow, J. Bartels, G. Clermont, and Y. Vodovotz, "A mathematical simulation of the inflammatory response to anthrax infection," *Shock*, vol. 29, pp. 104–111, 2008.

[50] A. Zelmer and T. H. Ward, "Noninvasive fluorescence imaging of small animals," *Journal of Microscopy*, vol. 252, pp. 8–15, 2013.

[51] L. Ottobrini, G. Lucignani, M. Clerici, and M. Rescigno, "Assessing cell trafficking by noninvasive imaging techniques: applications in experimental tumor immunology," *The Quarterly Journal of Nuclear Medicine and Molecular Imaging*, vol. 49, no. 4, pp. 361–366, 2005.

[52] E. J. Akins and P. Dubey, "Noninvasive imaging of cell-mediated therapy for treatment of cancer," *The Journal of Nuclear Medicine*, vol. 49, no. 6, pp. 180–195, 2008.

[53] Z. Han, A. Fu, H. Wang et al., "Noninvasive assessment of cancer response to therapy," *Nature Medicine—Technical Reports*, vol. 14, pp. 343–349, 2008.

[54] E. D. Nair-Gill, C. J. Shu, C. G. Radu, and O. N. Witte, "Non-invasive imaging of adaptive immunity using positron emission tomography," *Immunological Reviews*, vol. 221, no. 1, pp. 214–228, 2008.

[55] M. Chopra, S. Kraus, S. Schwinn et al., "Non-invasive bioluminescence imaging to monitor the immunological control of a plasmablastic lymphoma-like B cell neoplasia after hematopoietic cell transplantation," *PLoS ONE*, vol. 8, no. 12, Article ID e81320, 2013.

[56] N. G. B. Agrawal and J. J. Linderman, "Mathematical modeling of helper T lymphocyte/antigen-presenting cell interactions: analysis of methods for modifying antigen processing and presentation," *Journal of Theoretical Biology*, vol. 182, no. 4, pp. 1178–1190, 1996.

[57] Y. Vodovotz, G. Clermont, C. Chow, and G. An, "Mathematical models of the acute inflammatory response," *Current Opinion in Critical Care*, vol. 10, no. 5, pp. 383–390, 2004.

[58] F. Mitha, T. A. Lucas, F. Feng, T. B. Kepler, and C. Chan, "The Multiscale Systems Immunology project: software for cell-based immunological simulation," *Source Code for Biology and Medicine*, vol. 3, article 6, 2008.

[59] H. Miao, J. A. Hollenbaugh, M. S. Zand et al., "Quantifying the early immune response and adaptive immune response kinetics in mice infected with influenza A virus," *Journal of Virology*, vol. 84, no. 13, pp. 6687–6698, 2010.

[60] H.Y. Lee, D. J. Topham, S. Y. Park et al., "Simulation and prediction of the adaptive immune response to influenza A virus infection," *Journal of Virology*, vol. 83, no. 14, pp. 7151–7165, 2009.

[61] S. Marino and D. E. Kirschner, "The human immune response to Mycobacterium tuberculosis in lung and lymph node," *Journal of Theoretical Biology*, vol. 227, no. 4, pp. 463–486, 2004.

[62] S. D. Haessler and R. B. Brown, "Pneumonia caused by Staphylococcus aureus," *Current Respiratory Medicine Reviews*, vol. 5, no. 1, pp. 62–67, 2009.

[63] A. K. Abbas and A. H. Lichtman, *Basic Immunology Updated Edition: Functions and Disorders of the Immune System*, Elsevier Health Sciences, 2010.

[64] J. C. Strikwerda, *Finite difference schemes and partial differential equations*, Society for Industrial and Applied Mathematics, p, 2004.

[65] R. J. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations*, Society for Industrial and Applied Mathematics, Philadelphia, Pa, USA, 2007.

[66] H. C. Frey and S. R. Patil, "Identification and review of sensitivity analysis methods," *Risk Analysis*, vol. 22, no. 3, pp. 553–578, 2002.

[67] A. Saltelli, M. Ratto, S. Tarantola, and F. Campolongo, "Sensitivity analysis practices: strategies for model-based inference," *Reliability Engineering & System Safety*, vol. 91, no. 10-11, pp. 1109–1125, 2006.

[68] H. Wu, H. Miao, G. R. Warnes et al., "Dediscover : a computation and simulation tool for hiv viral fitness research," in *Proceedings of the International Conference on BioMedical Engineering and Informatics (BMEI '08)*, vol. 1, pp. 687–694, 2008.

[69] P. S. Kim, P. P. Lee, and D. Levy, "Modeling regulation mechanisms in the immune system," *Journal of Theoretical Biology*, vol. 246, no. 1, pp. 33–69, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 902545, 9 pages
http://dx.doi.org/10.1155/2014/902545

Review Article

Modeling Biology Spanning Different Scales: An Open Challenge

Filippo Castiglione,[1] Francesco Pappalardo,[2] Carlo Bianca,[3,4] Giulia Russo,[2] and Santo Motta[5]

[1] *Institute for Applied Mathematics, National Research Council of Italy, Rome, Italy*
[2] *Department of Pharmaceutical Sciences, University of Catania, Catania, Italy*
[3] *Theoretical Physics of Condensed Matter, Sorbonne Universities, UPMC Univ Paris 6, 75252 Paris Cedex 05, France*
[4] *UMR 7600 LPTMC, CNRS, 75252 Paris Cedex 05, France*
[5] *Department of Mathematics and Computer Science, University of Catania, 95125 Catania, Italy*

Correspondence should be addressed to Filippo Castiglione; f.castiglione@iac.cnr.it

Received 16 April 2014; Accepted 25 June 2014; Published 17 July 2014

Academic Editor: Vladimir Brusic

It is coming nowadays more clear that in order to obtain a unified description of the different mechanisms governing the behavior and causality relations among the various parts of a living system, the development of comprehensive computational and mathematical models at different space and time scales is required. This is one of the most formidable challenges of modern biology characterized by the availability of huge amount of high throughput measurements. In this paper we draw attention to the importance of multiscale modeling in the framework of studies of biological systems in general and of the immune system in particular.

1. Introduction

The language of mathematics has been extensively used to describe natural phenomena of the physical sciences in terms of models based on equations. The mathematical language allows logical reasoning over a representation of the physical entities involved in the phenomenon and makes possible to account for the observations made through experimentation.

In designing the mathematical model of a natural phenomenon the first and fundamental step is to define the mathematical variables that play a role in the phenomenon under investigations, according to the goals which the model is built for. For example, to calculate the decay rate of a certain protein, a variable to describe the changes of the protein concentration in the blood can be used. In this case the dynamics of the atoms and the ions is neglected and the information about the folding of the protein itself is lost. The origin of this oversight is related to the basic principle sometimes referred to as the *lex parsimoniae* most commonly known as the *Ockam's Razor*. "Pluralitas non est ponenda sine necessitate" in very simple words states that in the description of a phenomenon, the most useful model is the most parsimonious one in terms of elements used. In this regard, following up the above example, it makes little sense to describe the laws governing the forces accounting for the folding of the protein if we are interested in the half-life of the protein and we can estimate its decay rate by fitting a curve to a set of experimental data about the concentration in the blood of that protein.

William of Ockham was a Franciscan monk and logician who lived in the 14th century in a village of the English county of Surrey. At that time the principle of parsimony in describing and modeling a natural phenomenon was well reasoned. However, today the situation is a bit different. The *lex parsimoniae* is still valid and indeed is used when describing a phenomenon, but besides classical mathematical models allowing for an exact analytical approach, another modus operandi is now commonly employed [1–3]. This is what we can call the *synthetic approach* consisting in constructing a *replica or toy* of the studied system in terms of the most important identified elements and the laws governing the relationship among them. Actually this approach is not new

at all. The "engineer" Leonardo Da Vinci used this approach to construct toy models of flight machines before attempting anything real-scale.

What is new today is that we can use digital computers to construct toy models of complex systems. Indeed extremely powerful CPUs can be instructed to execute algorithms representing entities and laws and all kinds of conceptual experiments on those entities and laws can be made. This "digital synthetic" approach is commonly referred to as simulation.

Today, when studying a certain natural phenomena, scientists first identify elements and basic laws governing the dynamics of the system then they represent them as data structures and algorithms and finally execute the algorithms to observe how the system evolves. The Ockam's principle is still valid and used in the first phase of this process but beyond that the parsimony is forsaken, and the complexity of the initial toy model is augmented by simply adding new entities and laws. Indeed, with little difficulty we can detail processes incorporating hypothetical or experimentally derived knowledge. We can even *compose* preconstructed models of different parts of the real system or arrange models describing reality at different scales of observation, thus constructing a multiscale model. Ockam's Razor has been extensively used in classical mechanics models generating a cascade of models of increasing complexity. An interesting example arises from models of fluid dynamics which consider first uncompressible nonviscous fluids in a linear regime to move toward more complex situations like boundary layers and turbulent regimes. Models including different regimes are still difficult to perform.

This holistic approach is what in modern biology is called *systems biology* [4]. In this regard, there is another important aspect that should not be left out from the whole picture: the contemporary data explosion deriving from genomic, transcriptomics, proteomics, and metabolomics studies consisting in high dimensional datasets produced by latest high throughput measurements methods [5]. Other types of data coming from modern microscopy and biological imaging contribute as well to the detailed description of the constitutive parts and basic structures of living organisms [6]. On that account, the challenge has its main feature in relating these datasets to higher-level phenotypic characteristics and computational multiscale modeling approaches are set to reveal quantitative mechanistic relationships between these various measurements [7]. For example, high throughput gene expression data can be used to infer knowledge of the intracellular activities that can be later ascribed to the behavior of cells in a higher-level description; for example, the expression of the gene GATA3 in CD4 T lymphocytes in certain experimental conditions gives indication about the differentiation state of these cells and ultimately on the Th1 or Th2 bias of the immune response [8]. This information is relevant to the construction of a mathematical model of the immune response.

Recently, the topic of multiscale modeling has been drawn a great deal of attention and is discussed in many articles and reviews [6, 9–15]. Similarly, the present paper aims at giving a meaning to the concept of multiscale modeling in the framework of studies of biological systems in general but with particular interest in the immune system. It provides a general introduction to the methodological issues of multiscale modeling avoiding pointing to a specific and well-defined method to deal with this matter. Indeed, while there are methods borrowed from other field (e.g., computational chemistry) that can be used in special cases, a well-developed mathematical framework that is general enough to account for the extremely large variety of biological phenomena, is still missing. Nevertheless, an interesting attempt in this respect is given in [16] together with two examples showing how to bridge different single-scale models. Extensive readings, including specific examples, can be found in the above-cited reviews and also in [17–22].

It is worth stressing that the important role that the environment has in the dynamics of complex physics and living systems is not considered in this paper. Therefore the contents of the present refer to closed systems.

2. From Micro to Macro: Scales in Biological Organization

When "measuring" nature we choose a temporal and a spatial scale that is convenient to make a valid observation. The choice of the observation scale is an important step in science. In physics there is a somehow well-defined dividing line among different research areas based on the characteristic lengths of the systems studied and on the characteristic time of the phenomena under investigation. For instance, microphysics (e.g., molecular physics, atomic physics, nuclear physics, and particle physics) refers to areas of physics that study phenomena that take place at the microscopic scale (lengths < 1 mm). Similarly, in biology we can distinguish from molecular biology, microbiology, and cell biology looking at length scales below tenths of micrometers. Major levels of biological organization are regulated at scales of many orders of magnitude in space and time (see Figure 1), with space spanning from the molecular scale (10^{-10} m) to the living organism scale (1 m) and time from nanoseconds (10^{-9} s) to years (10^{8} s). In biology, while we can intuitively assert if a determined process involves cells, molecules, or organs, it is not so simple to identify values for the lengths at which we switch from one level to the next [6].

2.1. Single-Level Models. Roughly speaking, multiscale model is a composition of two or more "single" scale models representing the same phenomenon (or its parts) at different levels of descriptions. Even if the models we want to combine share the level of description, the manner in which the components are put together, namely, how the variables should be linked together, is a challenging part. For example, a simple model that describes the HIV infection of T helper lymphocytes may also take into account the coinfection of antigen presenting cells like macrophages and dendritic cells. Adding this new cell compartments to the original simplistic model introduces the problem of describing the immunological mechanisms of activation of the adaptive immunity by the innate one; in particular, the macrophages and the dendritic cells are both virus target and main actors of T helper priming.

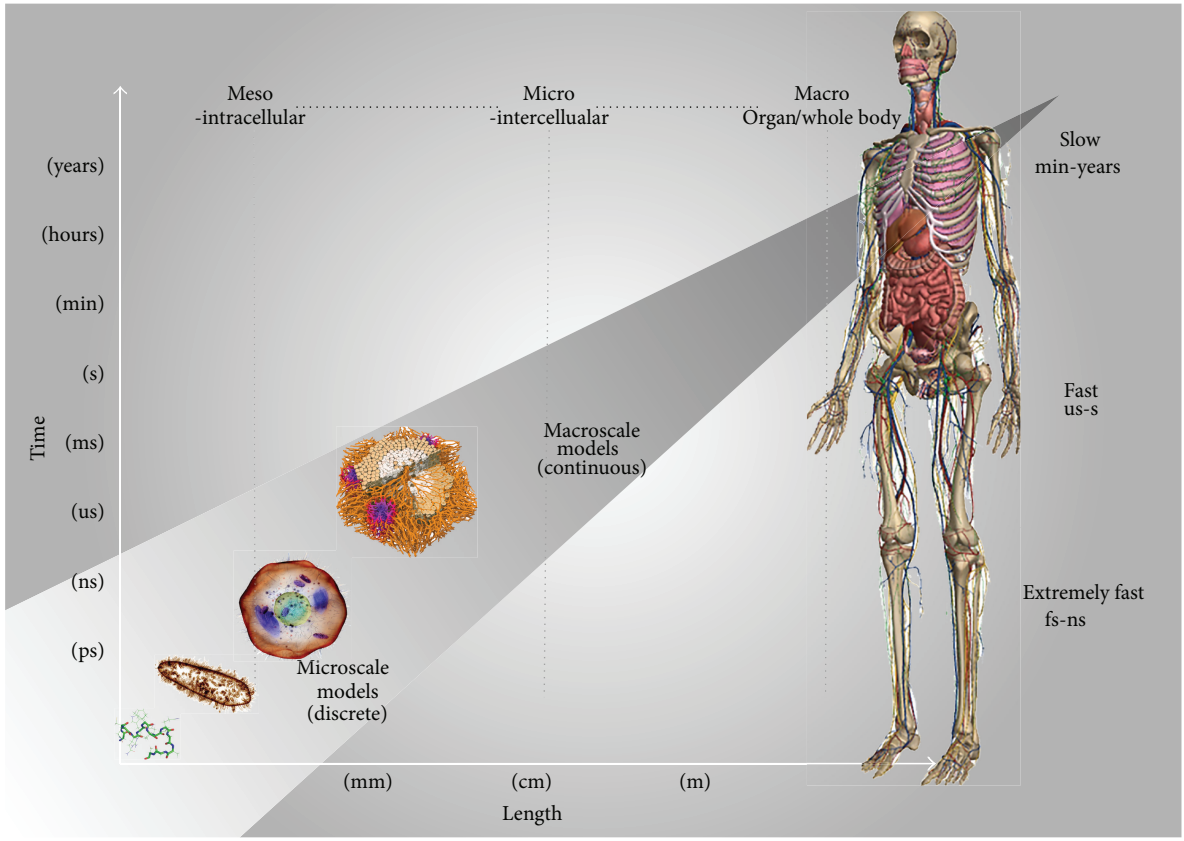


Figure 1: Multiscale models of the human body targeting complex processes span many time and length scales of biological organization. They cover a combination of discrete and continuous mathematical descriptions of different systemic components.

Moreover in biological phenomena complexity arises not only from the action of many independent actors, like in social science, but also from the fact that changes at lower scales modify the way in which those actors will play at higher scale. For instance DNA modification in a cell may change the cell in a tumor cell which then duplicates much faster than a normal one changing the overall scenario both at cellular and at tissue level. In most biological models this "vertical" or "interscale" complexity must be taken into account.

In the study of complex phenomena involving the immune system in pathological conditions, a unified view is necessary to reach a comprehension of the various mechanisms in action and of the causal relationships among different immune system components as well as repercussions on different anatomical parts [7]. More than for other complex systems, the distributed nature of immune system functions evidences the need of an integrated approach. The evolution of a disease like diabetes or cancer [23] is representative of this fact.

As already mentioned, mathematical models that try to describe such mechanisms, usually fix the spatial and temporal scale and describe the system with a mathematical or computational (i.e., algorithmic) formalism [11, 12, 14]. Computers do the rest as they provide the dynamics by executing (resolving) the rules just described in the mathematical formalism. The whole dynamics depends on parameters and initial conditions so that one generally attempts hypothetical scenarios by modifying those initial conditions to get a feeling of the systems behavior [20, 22]. This process leads itself in discovering new knowledge. However, the problem is that the real system is in general not isolated hence a local description is not sufficient to disclose crucial mechanisms. It comes quite clear that one of the reasons why biological phenomena are intrinsically complex is because they are influenced by variables that are outside a single level of space/temporal description. Moreover the collective behavior cannot be simply inferred from the behavior of its elements and the alteration of only one element or one interaction reverberates on the whole system. Finally a global organization emerges from the interacting elements (emergent behavior), which does not exist at the individual elements level.

3. Top-Down, Bottom-Up, or Middle-Out?

It should be noted that experiments are done at many scales, ranging from single molecules or proteins to whole organs and organisms, and therefore, experimental information exists at different scales. Therefore, relying on different experimental data, a model can be formulated using two main approaches, that is, top-down or bottom-up [24, 25]. If one chooses to take into account the individual elements and their interactions, studying the resulting biological system as a consequence of the emergent behavior of its single components, then the bottom-up approach takes place. For example one

can model the different immune system entities composed of cells and molecules to simulate the immune response against a specific pathogen, or one can use the cells as the basic elements and study tissue-level properties as results of the interactions of the cells. The advantage of this type of approach is that it is adaptive and robust, in the sense that if the available biological knowledge varies, one can adapt the new knowledge to the specific components of the model, in a very selective way. Moreover this kind of approach is suitable for studying the emergent properties of systems consisting of a large number of interacting elements. The intensive computer power required is the main disadvantage for the bottom-up approach and can be sometimes even prohibitive. Besides, the model itself can become too complicated to be controlled.

Instead, one can decide not to look straight into the details of the individual elements, but to consider the system at the macroscopic level, using experimental observations as guidelines during the formulation of the model. This is the case of the top-down modeling approach. For example, to keep on with the same example above, one can decide to model the immune system response against a specific pathogen ignoring the specific type of cells and their properties and modeling the global effect of population of cells, based on whole-cell experimental recordings. The clear advantage of this approach is that it is relatively simple. On the other hand, the flexibility and the robustness of the model are less evident compared with the bottom-up approach. Moreover, it should be highlighted that the variables and parameters in these models are largely phenomenological without direct connection with detailed physiological parameters. Due to this reason, it may sometimes happen that the top-down approach does not correctly reveal the actual responsible mechanism, for example, when there are multiple mechanisms for the same behavior or a single mechanism resulting in multiple effects. When existing components have to be integrated with some new part a third design principle, named "middle-out," is used [26]. This paradigm promotes the integration of organs' models at different scales without posing limitations to the level of details each single component should be equipped with (recall the *lex parsimoniae*). In this regard, it must be emphasized that multiscaling is not about sophistication but rather poses a different challenge, that of the model integration.

4. Multiscale Modeling of Biological Properties and Functions

Spanning from the lowest scale to higher levels, different modeling techniques can be chosen [27]. For intracellular scale, the modeling technique tries to give a detailed description of the molecular processes happening inside the cells. Using experimental data, these kinds of models make use of the differential equation description to forecast the molecular dynamics of specific cellular pathways. Changes in the molecular concentrations are described by these models by mass action or Michaelis-Menten kinetic rate-law equations.

The Belousov-Zhabotinsky reaction represents a good example of a bidomain model that depicts a phenomenon beginning from the microscopic dynamics at a lower space scale, that is, wave propagation in reactive media. In its simple form it may be comprehended in terms of the following representation [28] including an autocatalytic reaction $A + Y \rightarrow X + P$, $X + Y \rightarrow 2P$, $A + X \rightarrow 2X + 2Z$, $2X \rightarrow A + P$, and $B + Z \rightarrow hY + Q$, where the variables represent concentrations of specific molecules (e.g., bromomalonic acid or carbon dioxide) and h is a constant. Translated to ordinary differential equation the system is $dX/dt = AY - XY + AX - 2X^2$, $dY/dt = -AY - XY + hBZ$, and $dZ/dt = 2AX - BZ$, where A, B, and P are held constant. The multiscale property of this model is found in the occurrence of the wave at a level that is above the one chosen to describe the phenomena, that is, the molecular level of the reactants.

The main difficulty is represented by parameter identification: the experimental estimation is often made in isolated systems that, by definition, do not permit generalization to the real case. If the interacting entities in a system to be modeled can be thought as homogeneous, then the most common choice is the use of ordinary differential equations. If the space is variable, then partial differential equations can represent a better technique [29].

In the case of intracellular models that consider small number of entities, microsimulation can represent an alternative to differential equations. The authors in [30, 31] proposed the Gillespie algorithm many decades ago. It allows simulating with a good accuracy chemical or biochemical systems of reactions generating statistically correct trajectories as possible solutions of a stochastic equation.

At a higher level of description, tissues or whole organs are modeled in two different ways: either as functional compartments or system units or as a collection of microscopic components (e.g., cells). In the first case rather than specifically model the organ, one can simply use the known input-output relationship as a black box. This relation is typically derived from experimental data or published results and ultimately developed by differential equations. These kind of phenomenological models aim at reproducing the observed behavior instead of trying to give an explanation. The modeling paradigm based on a collection of microscopic components intends to typify a tissue as an array of individual units (i.e., cells) exchanging signals with the environment. Examples of these multicellular systems have been originally developed to study the growth of solid tumors [32, 33] and have later on been applied to simulate the function (the regeneration) of complex organs like the liver [34].

An interesting example of a well-devised multiscale model has been developed in the framework of the hemodynamics [35, 36]. The problem deals with a detailed description of the fluid-dynamics of the blood, by mean of numerical integration of the Navier-Stokes equations, to cope with postoperative hemodynamics issues in congenital heart diseases, artery shunts, or similar heart surgery. In hemodynamics, local phenomena, such as the perturbation of flow pattern in a specific vascular region, are strictly related to the global features of the whole circulation. However, dealing with whole circulation using Navier-Stokes equations would be not just useless but rather impossible.

The method proposes an interesting multiscale geometrical model where a local, accurate, three-dimensional description of blood flow by means of the Navier-Stokes equations in a specific artery (the region of surgical interest) is coupled with a systemic, zero-dimensional, lumped model of the rest of the circulation system [37]. What makes the example peculiar is the use of lumped models as those extensively used in electrical engineering that resort to simplified models in place of complex description of other system parts.

Another methodology worth to be mentioned is the one using "state transition diagram" [38, 39] which aims to solve the problem of heterogeneity and multiscale modeling and the link between mathematical and computer models [40]. This methodology, massively used in theoretical computer science and software engineering, describes the behavior of heterogeneous entities by means of (deterministic or probabilistic) finite state automata. Since each state of the automaton represents a "situation" related to a level of description, one could in principle set out a multiscale model as a combination of automata. However, since the number of states resulting from the combination of even simple models tends to be very large, this methodology does not seem to be of practical use.

If the interest is on simulating a whole cell, then several projects can provide useful hints (e.g., virtual cell [41], e-cell [42, 43]), whereas efforts aiming at simulating whole physiological systems or organs are, for example, models of the heart [18], of the liver [44], and of the skeletal system [45]. Other efforts aim at creating computational platforms suite to integrating various physiological processes by integration of different mathematical and computational models [46]. The approach is based on the principle that, in biology, there is no privileged level for the description of a certain phenomenon and that the interlevel causal relationships are driven by interactions between multiple levels [47–50].

It is worth stressing that the modelling of complex biological systems requires a completely different treatment with respect to the inert matter. Indeed the entities constituting the biological systems, which usually operate out-of-equilibrium, interact among themselves and with their outer environment and are able to perform individual strategies that modify the microscopic interactions among the entities composing the system [51].

Recently the kinetic theory has proposed an alternative approach for deriving macroscopic equations from the dynamics delivered at the mesoscopic scale: the asymptotic method. Accordingly, this method consists in deriving macroscopic equations by suitable limits of Boltzmann-type equations related to the statistical microscopic description; see the book [52], the paper [53], and the references cited therein. The first step in the development of asymptotic methods is the choice of the time-space scaling. Different types of scaling lead to different types of equations. After the assessment of the scaling, the distribution function is expanded in terms of a small dimensionless parameter. Finally the asymptotic limit is performed under suitable technical assumptions. Specifically parabolic (or low-field) scaling of kinetic equations leads to a drift-diffusion type macroscopic system where the diffusion processes dominate the behavior of the solutions.

In the hyperbolic (or high-field) limit the influence of the diffusion terms is of lower (or equal) order of magnitude in comparison with other convective or interaction terms and the models consist of linear or nonlinear hyperbolic equations for the local density.

Finally the use of kinetic models coupled with deterministic thermostats has been recently proposed for the modeling of complex biological systems subjected to external force field, such as a vaccine, but constrained to keep constant the total energy; see [54].

5. Multiscale Methods

From the computational point of view, there are methods employed in other field of science that can potentially be employed in biology [11]. These are the Quasi-continuum, the hybrid quantum mechanics-molecular mechanics, the equation-free, the heterogeneous, the multigrid, the multiscale agent-based modeling, the multiscale numerical scheme, and the adaptive tabulation approach. Although we do not describe them here for brevity (suggesting reference [11] as a good starting point), we care to say that despite the fact that each of these has been efficiently applied in a specific problem domain and each has its pro and contra in terms of computational efficiency, none of them has emerged as the multimethod to be used to model biological phenomena.

One example of multiscale approach we care to give more details on is the one we have used to set up a model of (type I) hypersensitive phenomena. According to what just said, it can be classified as a multiscale agent-based model. It consists in an agent-based formulation of the cell-cell/molecules interaction pertaining to hypersensitive responses to a generic allergen in which a detailed gene regulation dynamics is modeled by means of a Boolean network [55] (other approaches, as the use of a system of ordinary differential equations, would work as well [56]). The two levels (the intra- and the inter-cellular) are integrated in a quite intuitive way. For each T lymphocyte, the intracellular gene regulation is driven by the extracellular cytokine concentration (consider it as the *cell input*). On the other hand, the transcription of certain genes can drive the differentiation of the cell and also the production of other cytokines (i.e., the *output*) which influences the overall immune dynamics [8].

What makes this approach appealing is that omics data can effectively be integrated with cellular level data largely available, making a genetic-cause/phenotypic-effect analysis possible [57]. The kind of information clinicians is looking after. Moreover, the two levels of descriptions (the gene regulation through networks) and the intercellular dynamics of the immune response can be developed independently one from another and later put together to account for a more elaborate description of the same, or of other, phenomena. An example would be a detailed description of T helper differentiation in four phenotypes (Th1, Th2, T regulatory, and Th17) [58] which is at the core of, for example, inflammation phenomena, with an agent-based simulation of the immune

response not just of infective pathogens or allergens, but also of inflammation and emergence of type 2 diabetes [59, 60].

Other works also incorporated networks or ODEs in agent-based models. See for example, [61] in which the authors discuss the combination of ODEs for chemokine receptor internalisation with agent-based models of lymphocytes in the context of tissue instability in arthritis. Also in [62] the authors describe an approach in which they combine the molecular, the cellular, and the tissue scale in a spatial model of the intestinal crypt. Moreover in [63] Perfahl et al. discuss the domain size effects in the context of vascular tumors in a 3D agent-based approach combined with a reaction-diffusion system.

Kirschner et al. have provided different examples of multiscale immune simulation combining the agent-based paradigm to represent one level of description (i.e., the cellular mesoscopic level) combined to ordinary differential equations. In [64] the authors describe the immune response to M. Tuberculosis representing cells as agents and describing the time-dependent processes essential to antigen processing and presentation by means of ordinary differential equations.

In another work [65], the same authors present an approach for integrating information over relevant biological and temporal scales to generate such a representation for major histocompatibility complex class II-mediated antigen presentation. They then show how this kind of models can be used to suggest new mechanisms and strategies for treatment and vaccines.

When both stochastic fluctuations and spatial inhomogeneity must be included in a model simultaneously, the resulting computational demand quickly becomes overwhelming. In this case it would be useful to use an approach based on coarse-graining methods which turns out to be essential for realistic multiscale models. For instance in [66], the authors present an algorithm for simulation of stochastic, spatially inhomogeneous reaction-diffusion kinetics coupled to coarse-grained fields described by (stochastic or deterministic) partial differential equations (PDEs). They successfully used this method to model cell signaling dynamics in spatially inhomogeneous environments and under the influence of external fields.

5.1. General Purpose Integration Methods. When developing a multiscale approach there are few aspects that need to be taken into account. In general, the time scales on which the lower-level processes occur are much faster than those on which the higher-level processes occur. Usually the lower-level processes can be assumed to occur instantaneously and can therefore be included as a representation of some kind of field at the higher level [6]. When we consider joining independent models of processes that occur on different scales, it is enticing to simply couple existing components (i.e., software) for the separate models to one another. This way to proceed does not consider how inaccuracies in the values of the variables that are passed between the two models may affect the combined model. In order to prevent these inaccuracies from occurring one should consider the whole as a single model rather than the combination of two simpler ones. For instance, we can consider that a microscopic simulator at the cellular level can be coupled with the description of the intracellular signaling activating a specific cellular pathway. In this example the differentiation of T lymphocytes into the phenotypes Th1, Th2, Treg, and Th17 is described at a cellular level by means of individual entities (e.g., agent-based) whereas the gene regulation is described by a system of differential equations describing activation level of each gene of the gene network represented with the following equation: $dx_i/dt = (-e^{-Ch} + e^{-h(\omega_i - C)})/((1-e^{Ch})(1+e^{-h(\omega_i - C)})) - \gamma_i x_i$, where t is the time, x_i is the activation level of the ith gene, ω_i and γ_i are parameters relative to the network topology, and C and h are constants [55]. Here the lower level description of gene activation is determined at each upper-level time step by solving the system of ODEs and the cell differentiation is executed at the upper level on the basis of the information coming from the gene expression levels. This procedure is iteratively executed at each time step and for each lymphocyte.

From a computational perspective the multiscale nature of innovative models has prompted the important issue of reusability of available published models targeting a single scale. The Physiome project [17, 23] is a prominent effort aiming at solving this problem by developing a framework for the modeling of the "whole" human body. As part of that initiative, the markup language CellML was introduced with the aim of establishing a worldwide adopted standard in the development of cellular levels that are modeled as sets of ODEs [67]. Similarly, FieldML has been defined to model processes on the tissue and organ level that are represented as sets of PDEs [68]. Systems biology markup language (SBML) [69] has been proposed and is now beginning to make a significant impact on the modeling community as a means to exchange models. However, neither CellML nor SBML includes explicit directives to deal with the problem of implementing a multiscale computational model. To solve this important issue, however, there are some attempts, for example, MML [70].

A framework that is devoted to the systems biology community with the target of easy model interoperability is represented by the systems biology workbench [71], a high performance, open-source software infrastructure that allows heterogeneous application components written in diverse programming languages and running on different platforms to communicate and use each others' capabilities via a message system.

6. Concluding Remarks

In the study of complex biological phenomena it is necessary to develop a unified view of the various mechanisms in action and of the causal relationships among different parts of that complex system, [4, 7]. In this paper we have briefly described the problems faced when one wants to link mathematical or computational models across different time and length scales.

In many areas of biology and physiology, multiscale and multiphysics models are very much acclaimed, although there exists an abundant literature for multiscale models in science and engineering domains [72], a lot remains to be

done in terms of translating these mathematical theories and methodologies to the domains of biology and physiology [73–75].

A key unsolved issue is how to represent appropriately the dynamical behaviors of a high-dimensional model of a lower scale by a low-dimensional model of a higher scale, so that it can be used to investigate complex dynamical behaviors at even higher scales of integration [14]. Indeed, the ultimate goal of multiscale modeling is not just about developing models at different scales but to link them in a consistent manner so that the information from a lower scale can be carried into the simplified model of a higher scale.

The use of different modeling approaches introduces gaps among scales. Multiscale modeling, besides modeling the system, needs to address the issue of how to bridge the gaps between different methodologies and between models at different scales. Unfortunately, there is no specific or simple way to tell how to achieve this objective, but there are empirical principles and methods that can be of help.

In the study of the immune system and related pathologies, one method for constructing multiscale models that has been used by various authors resorts to agents to represent the mesoscopic level of cells of the immune system (i.e., the multicellular rule-based modeling in [76]) while employing ordinary differential equations to describe the intracellular events as intracellular signalling and partial differential equations to describe cytokines diffusion at the extracellular or tissue scale. Level coupling is then performed in a quite straightforward way using concentrations as input variables to the cellular agents. Whereas modeling intracellular events can be implemented in many ways (e.g., Boolean networks or other generic decision mechanisms) without explicitly including the variable "space" for computational reasons (but mainly for simplicity), the diffusion of cytokines (or, another example, cells relocation between anatomical compartments), is a spatial phenomenon in character. This can be modeled as a continuous (by means of PDEs) or as a discrete process (e.g., lattice gas) for which the computational efficiency is the major limiting factor.

The goal of computational systems biology is to consider a biological system from a holistic perspective and use both experiments and modeling to reveal how the system behaves [4, 77]. Multiscale models able to exploit laboratory and clinical data at different levels can potentially bridge knowledge gaps between what is observed at the gene/molecular level and the clinical evolution of complex diseases [11].

Finally, by integrating these models with detailed monitoring data from emerging body-sensor technology [78], health care practitioners could be supported in taking diagnosis and suggesting optimal therapeutic regimens thus promoting the much acclaimed patient-specific view of modern health care systems.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Authors' Contribution

Filippo Castiglione and Francesco Pappalardo equally contributed to the work.

Acknowledgments

Filippo Castiglione acknowledges partial support from the European Commission under the 7th Framework Programme (MISSION-T2D Project, Contract no. 600803). Santo Motta acknowledges partial support from PRIN 2009, "Metodi e Modelli Matematici della Teoria Cinetica per Sistemi Complessi." Carlo Bianca acknowledges partial support from the L' Agence Nationale de la Recherche (ANR T-KiNeT Project).

References

[1] A. Baker, "Simplicity," in *Stanford Encyclopedia of Philosophy*, Stanford University, Stanford, Calif, USA, 2010.

[2] N. Oreskes, K. Shrader-Frechette, and K. Belitz, "Verification, validation, and confirmation of numerical models in the earth sciences," *Science*, vol. 263, no. 5147, pp. 641–646, 1994.

[3] E. Sober, "Let's razor Occam's razor," in *Explanation and Its Limits*, D. Knowles, Ed., pp. 73–93, Cambridge University Press, 1991.

[4] H. Kitano, "Systems biology: a brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[5] C. M. Deane, Ł. Salwiński, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations," *Molecular & Cellular Proteomics*, vol. 1, no. 5, pp. 349–356, 2002.

[6] J. Southern, J. Pitt-Francis, J. Whiteley et al., "Multi-scale computational modelling in biology and physiology," *Progress in Biophysics and Molecular Biology*, vol. 96, no. 1–3, pp. 60–89, 2008.

[7] B. di Ventura, C. Lemerle, K. Michalodimitrakis, and L. Serrano, "From *in vivo* to *in silico* biology and back," *Nature*, vol. 443, no. 7111, pp. 527–533, 2006.

[8] D. Santoni, M. Pedicini, and F. Castiglione, "Implementation of a regulatory gene network to simulate the TH1/2 differentiation in an agent-based model of hypersensitivity reactions," *Bioinformatics*, vol. 24, no. 11, pp. 1374–1380, 2008.

[9] J. B. Bassingthwaighte, H. J. Chizeck, L. E. Atlas, and H. Qian, "Multiscale modeling of cardiac cellular energetics," *Annals of the New York Academy of Sciences*, vol. 1047, pp. 395–424, 2005.

[10] P. V. Coveney and P. W. Fowler, "Modelling biological complexity: a physical scientist's perspective," *Journal of the Royal Society Interface*, vol. 2, no. 4, pp. 267–280, 2005.

[11] J. O. Dada and P. Mendes, "Multi-scale modelling and simulation in systems biology," *Integrative Biology*, vol. 3, no. 2, pp. 86–96, 2011.

[12] A. J. Engler, P. O. Humbert, B. Wehrle-Haller, and V. M. Weaver, "Multiscale modeling of form and function," *Science*, vol. 324, no. 5924, pp. 208–212, 2009.

[13] R. Grima, "Multiscale modeling of biological pattern formation," *Current Topics in Developmental Biology*, vol. 81, pp. 435–460, 2008.

[14] Z. Qu, A. Garfinkel, J. N. Weiss, and M. Nivala, "Multi-scale modeling in biology: how to bridge the gaps between scales?"

Progress in Biophysics and Molecular Biology, vol. 107, no. 1, pp. 21–31, 2011.

[15] C. Lavelle, H. Berry, G. Beslon et al., "From molecules to organisms: towards multiscale integrated models of biological systems," *Theoretical Biology Insights*, vol. 1, pp. 13–22, 2008.

[16] P. M. A. Sloot and A. G. Hoekstra, "Multi-scale modelling in computational biomedicine," *Briefings in Bioinformatics*, vol. 11, no. 1, Article ID bbp038, pp. 142–152, 2010.

[17] C. Bradley, A. Bowery, R. Britten et al., "OpenCMISS: a multi-physics & multi-scale computational infrastructure for the VPH/Physiome project," *Progress in Biophysics and Molecular Biology*, vol. 107, no. 1, pp. 32–47, 2011.

[18] P. Hunter and P. Nielsen, "A strategy for integrative computational physiology," *Physiology*, vol. 20, no. 5, pp. 316–325, 2005.

[19] H. Joshi, A. B. Singharoy, Y. V. Sereda, S. C. Cheluvaraja, and P. J. Ortoleva, "Multiscale simulation of microbe structure and dynamics," *Progress in Biophysics and Molecular Biology*, vol. 107, no. 1, pp. 200–217, 2011.

[20] M. Meier-Schellersheim, I. D. Fraser, and F. Klauschen, "Multiscale modeling for biologists," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 1, no. 1, pp. 4–14, 2009.

[21] T. Murtola, A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen, "Multiscale modeling of emergent materials: biological and soft matter," *Physical Chemistry Chemical Physics*, vol. 11, no. 12, pp. 1869–1892, 2009.

[22] S. Schnell, R. Grima, and P. K. Maini, "Multiscale modeling in biology," *The American Scientist*, vol. 95, no. 2, pp. 134–142, 2007.

[23] P. J. Hunter and T. K. Borg, "Integration from proteins to organs: the physiome project," *Nature Reviews Molecular Cell Biology*, vol. 4, no. 3, pp. 237–243, 2003.

[24] L. Alberghina and H. V. Westerhoof, Eds., *Systems Biology—Definitions and Perspectives*, Springer, Heidelberg, Germany, 2008.

[25] C. Bianca and M. Pennisi, "Immune systems modelling by top-down and bottom-up approaches," *International Mathematical Forum*, vol. 7, no. 1–4, pp. 109–128, 2012.

[26] P. J. Hunter and M. Viceconti, "The VPH-Physiome project: Standards and tools for multiscale modeling in clinical applications," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 40–53, 2009.

[27] S. Motta and F. Pappalardo, "Mathematical modeling of biological systems," *Briefings in Bioinformatics*, vol. 14, no. 4, pp. 411–422, 2013.

[28] J. T. Tyson, "What everyone should know about the Belousov-Zhabotinsky reaction," in *Frontiers in Mathematical Biology*, S. A. Levin, Ed., pp. 569–587, Springer, New York, NY, USA, 1994.

[29] J. D. Murray, *Mathematical Biology Vol I and Vol II*, Springer, New York, NY, USA, 2003.

[30] D. T. Gillepsie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of Computational Physics*, vol. 22, no. 4, pp. 403–434, 1976.

[31] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[32] D. Drasdo, R. Kree, and J. S. McCaskill, "Monte Carlo approach to tissue-cell populations," *Physical Review E*, vol. 52, no. 6, pp. 6635–6657, 1995.

[33] D. Drasdo, "Buckling instabilities of one-layered growing tissues," *Physical Review Letters*, vol. 84, no. 18, pp. 4244–4247, 2000.

[34] S. Hoehme, M. Brulport, A. Bauer et al., "Prediction and validation of cell alignment along microvessels as order principle to restore tissue architecture in liver regeneration," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 23, pp. 10371–10376, 2010.

[35] G. Pennati, F. Migliavacca, G. Dubini, R. Pietrabissa, and M. R. De Leval, "A mathematical model of circulation in the presence of the bidirectional cavopulmonary anastomosis in children with a univentricular heart," *Medical Engineering and Physics*, vol. 19, no. 3, pp. 223–234, 1997.

[36] G. Pennati, M. Bellotti, and R. Fumero, "Mathematical modelling of the human foetal cardiovascular system based on Doppler ultrasound data," *Medical Engineering and Physics*, vol. 19, no. 4, pp. 327–335, 1997.

[37] K. Laganà, G. Dubini, F. Migliavacca et al., "Multiscale modelling as a tool to prescribe realistic boundary conditions for the study of surgical procedures," *Biorheology*, vol. 39, no. 3-4, pp. 359–364, 2002.

[38] O. Vainas, D. Harel, I. R. Cohen, and S. Efron, "Reactive animation: from piecemeal experimentation to reactive biological systems," *Autoimmunity*, vol. 44, no. 4, pp. 271–281, 2011.

[39] H. Bersini, D. Klatzmann, A. Six, and V. Thomas-Vaslin, "State-transition diagrams for biologists," *PLoS ONE*, vol. 7, no. 7, Article ID e41165, 2012.

[40] C. H. McEwan, H. Bersini, D. Klatzmann, V. Thomas-Vaslin, and A. Six, "Refitting harel statecharts for systemic mathematical models in computational immunology," in *Artificial Immune Systems*, vol. 6825 of *Lecture Notes in Computer Science*, pp. 44–50, Springer, Berlin, Germany, 2011.

[41] J. Schaff, C. C. Fink, B. Slepchenko, J. H. Carson, and L. M. Loew, "A general computational framework for modeling cellular structure and function," *Biophysical Journal*, vol. 73, no. 3, pp. 1135–1146, 1997.

[42] D. Normile, "Building working cells "in silico"," *Science*, vol. 284, no. 5411, pp. 80–81, 1999.

[43] K. Takahashi, K. Kaizu, B. Hu, and M. Tomita, "A multi-algorithm, multi-timescale method for cell simulation," *Bioinformatics*, vol. 20, no. 4, pp. 538–546, 2004.

[44] H-G. Holzhütter, D. Drasdo, T. Preusser et al., "The virtual liver: a multidisciplinary, multilevel challenge for systems biology," *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, vol. 4, no. 3, pp. 221–235, 2012.

[45] M. Viceconti, *Multiscale Modeling of the Skeletal System*, Cambridge University Press, New York, NY, USA, 2012.

[46] T. Eissing, L. Kuepfer, C. Becker et al., "A computational systems biology software platform for multiscale modeling and simulation: integrating whole-body physiology, disease biology, and molecular reaction networks," *Frontiers in Physiology*, vol. 2, article 4, 2011.

[47] S. Brenner, "Biological computation," in *The Limits of Reductionism in Biology*, G. Bock and J. Goode, Eds., vol. 213 of *Novartis Foundation Symposium*, pp. 106–116, John Wiley & Sons, London, UK, 1998.

[48] D. Noble, "Modeling the heart—from genes to cells to the whole organ," *Science*, vol. 295, no. 5560, pp. 1678–1682, 2002.

[49] D. Noble, *The Music of Life. Biology beyond the Genome*, Oxford University Press, Oxford, UK, 2006.

[50] A. A. de Graaf, A. P. Freidig, B. de Roos et al., "Nutritional systems biology modeling: from molecular mechanisms to physiology," *PLoS Computational Biology*, vol. 5, no. 11, Article ID e1000554, 2009.

[51] C. Bianca, "Thermostatted models—multiscale analysis and tuning with real-world systems data," *Physics of Life Reviews*, vol. 9, no. 4, pp. 418–425, 2012.

[52] C. Bianca and N. Bellomo, "Multiscale modeling: linking molecular, cellular, and tissues scales," in *Towards a Mathematical Theory of Complex Biological Systems*, vol. 11, pp. 89–115, 2011.

[53] A. Bellouquid and C. Bianca, "Modelling aggregation-fragmentation phenomena from kinetic to macroscopic scales," *Mathematical and Computer Modelling*, vol. 52, no. 5-6, pp. 802–813, 2010.

[54] C. Bianca, "Thermostatted kinetic equations as models for complex systems in physics and life sciences," *Physics of Life Reviews*, vol. 9, no. 4, pp. 359–399, 2012.

[55] L. Mendoza and F. Pardo, "A robust model to describe the differentiation of T-helper cells," *Theory in Biosciences*, vol. 129, no. 4, pp. 283–293, 2010.

[56] P. Martínez-Sosa and L. Mendoza, "The regulatory network that controls the differentiation of T lymphocytes," *BioSystems*, vol. 113, no. 2, pp. 96–103, 2013.

[57] M. Pedicini, F. Barrenäs, T. Clancy et al., "Combining network modeling and gene expression microarray analysis to explore the dynamics of Th1 and Th2 cell regulation," *PLoS Computational Biology*, vol. 6, no. 12, Article ID e1001032, 8 pages, 2010.

[58] A. Naldi, J. Carneiro, C. Chaouiya, and D. Thieffry, "Diversity and plasticity of Th cell types predicted from regulatory network modelling," *PLoS Computational Biology*, vol. 6, no. 9, Article ID e1000912, 2010.

[59] F. Castiglione, P. Tieri, A. de Graaf et al., "The onset of type 2 diabetes: proposal for a multi-scale model," *JMIR Research Protocols*, vol. 2, no. 2, p. e44, 2013.

[60] C. Bianca and A. Lemarchand, "Density evolution by the low-field limit of kinetic frameworks with thermostat and mutations," *Communications in Nonlinear Science and Numerical Simulation*, 2014.

[61] T. Beyer and M. Meyer-Hermann, "Cell transmembrane receptors determine tissue pattern stability," *Physical Review Letters*, vol. 101, no. 14, Article ID 148102, 2008.

[62] P. Buske, J. Galle, N. Barker, G. Aust, H. Clevers, and M. Loeffler, "A comprehensive model of the spatio-temporal stem cell and tissue organisation in the intestinal crypt," *PLoS Computational Biology*, vol. 7, no. 1, Article ID e1001045, 2011.

[63] H. Perfahl, H. M. Byrne, T. Chen et al., "Multiscale modelling of vascular tumour growth in 3D: The roles of domain size and boundary conditions," *PLoS ONE*, vol. 6, no. 4, Article ID e14790, 2011.

[64] D. Kirschner, "The multi-scale immune response to pathogens: *M. tuberculosis* as an example," in *In Silico Immunology*, D. Flower and J. Timmis, Eds., pp. 289–311, Springer, Berlin, Germany, 2007.

[65] D. E. Kirschner, S. T. Chang, T. W. Riggs, N. Perry, and J. J. Linderman, "Toward a multiscale model of antigen presentation in immunity," *Immunological Reviews*, vol. 216, no. 1, pp. 93–118, 2007.

[66] D. C. Wylie, Y. Hori, A. R. Dinner, and A. K. Chakraborty, "A hybrid deterministic-stochastic algorithm for modeling cell signaling dynamics in spatially inhomogeneous environments and under the influence of external fields," *Journal of Physical Chemistry B*, vol. 110, no. 25, pp. 12749–12765, 2006.

[67] A. Garny, D. P. Nickerson, J. Cooper et al., "CellML and associated tools and techniques," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1878, pp. 3017–3043, 2008.

[68] G. R. Christie, P. M. F. Nielsen, S. A. Blackett, C. P. Bradley, and P. J. Hunter, "FieldML: concepts and implementation," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1895, pp. 1869–1884, 2009.

[69] M. Hucka, A. Finney, H. M. Sauro et al., "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

[70] J.-L. Falcone, B. Chopard, and A. Hoekstra, "MML: towards a multiscale modeling language," *Procedia Computer Science*, vol. 1, no. 1, pp. 819–826, 2010.

[71] H. M. Sauro, M. Hucka, A. Finney et al., "Next Generation Simulation Tools: The Systems Biology Workbench and BioSPICE Integration," *OMICS A Journal of Integrative Biology*, vol. 7, no. 4, pp. 355–372, 2003.

[72] J. Fish, Ed., *Multiscale Methods, Bridging the Scales in Science and Engineering*, Oxford University Press, 2009.

[73] D. J. W. Evans, P. V. Lawford, J. Gunn et al., "The application of multiscale modelling to the process of development and prevention of stenosis in a stented coronary artery," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1879, pp. 3343–3360, 2008.

[74] A. Caiazzo, D. Evans, J. Falcone et al., "A complex automata approach for in-stent restenosis: two-dimensional multiscale modelling and simulations," *Journal of Computational Science*, vol. 2, no. 1, pp. 9–17, 2011.

[75] H. Tahir, A. G. Hoekstra, E. Lorenz et al., "Multi-scale simulations of the dynamics of in-stent restenosis: impact of stent deployment and design," *Interface Focus*, vol. 1, no. 3, pp. 365–373, 2011.

[76] A. K. Chavali, E. P. Gianchandani, K. S. Tung, M. B. Lawrence, S. M. Peirce, and J. A. Papin, "Characterizing emergent properties of immunological systems with multi-cellular rule-based computational modeling," *Trends in Immunology*, vol. 29, no. 12, pp. 589–599, 2008.

[77] P. Kohl, E. J. Crampin, T. A. Quinn, and D. Noble, "Systems biology: an approach," *Clinical Pharmacology and Therapeutics*, vol. 88, no. 1, pp. 25–33, 2010.

[78] F. Castiglione, V. Diaz, A. Gaggioli et al., "Physio-environmental sensing and live modeling," *Journal of Medical Internet Research*, vol. 15, no. 1, article e3, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 437987, 9 pages
http://dx.doi.org/10.1155/2014/437987

Research Article

Big Data Analytics in Immunology: A Knowledge-Based Approach

Guang Lan Zhang,[1] Jing Sun,[2] Lou Chitkushev,[1] and Vladimir Brusic[1]

[1] *Department of Computer Science, Metropolitan College, Boston University, Boston, MA 02215, USA*
[2] *Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA*

Correspondence should be addressed to Vladimir Brusic; vbrusic@bu.edu

Received 31 March 2014; Accepted 7 May 2014; Published 22 June 2014

Academic Editor: Francesco Pappalardo

With the vast amount of immunological data available, immunology research is entering the big data era. These data vary in granularity, quality, and complexity and are stored in various formats, including publications, technical reports, and databases. The challenge is to make the transition from data to actionable knowledge and wisdom and bridge the knowledge gap and application gap. We report a knowledge-based approach based on a framework called KB-builder that facilitates data mining by enabling fast development and deployment of web-accessible immunological data knowledge warehouses. Immunological knowledge discovery relies heavily on both the availability of accurate, up-to-date, and well-organized data and the proper analytics tools. We propose the use of knowledge-based approaches by developing knowledgebases combining well-annotated data with specialized analytical tools and integrating them into analytical workflow. A set of well-defined workflow types with rich summarization and visualization capacity facilitates the transformation from data to critical information and knowledge. By using KB-builder, we enabled streamlining of normally time-consuming processes of database development. The knowledgebases built using KB-builder will speed up rational vaccine design by providing accurate and well-annotated data coupled with tailored computational analysis tools and workflow.

1. Introduction

Data represent the lowest level of abstraction and do not have meaning by themselves. Information is data that has been processed so that it gives answers to simple questions, such as "what," "where," and "when." Knowledge represents the application of data and information at a higher level of abstraction, a combination of rules, relationships, ideas, and experiences, and gives answers to "how" or "why" questions. Wisdom is achieved when the acquired knowledge is applied to offer solutions to practical problems. The data, information, knowledge, and wisdom (DIKW) hierarchy summarizes the relationships between these levels, with data at its base and wisdom at its apex and each level of the hierarchy being an essential precursor to the levels above (Figure 1(a)) [1, 2]. The acquisition cost is lowest for data acquisition and highest for knowledge and wisdom acquisition (Figure 1(b)).

In immunology, for example, a newly sequenced molecular sequence without functional annotation is a data point, information is gained by annotating the sequence to answer questions such as which viral strain it originates from, knowledge may be obtained by identifying immune epitopes in the viral sequence, and the design of a peptide-based vaccine using the epitopes represents the wisdom level. Overwhelmed by the vast amount of immunological data, to make the transition from data to actionable knowledge and wisdom and bridge the knowledge gap and application gap, we are confronted with several challenges. These include asking the "right questions," handling unstructured data, data quality control (garbage in, garbage out), integrating data from various sources in various formats, and developing specialized analytics tools with the capacity to handle large volume of data.

The human immune system is a complex system comprising the innate immune system and the adaptive immune system. There are two branches of adaptive immunity, humoral immunity effected by the antibodies and cell-mediated immunity effected by the T cells of the immune

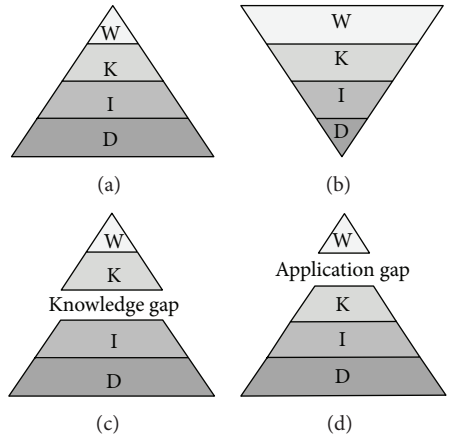


Figure 1: The DIKW hierarchy. (a) The relative quantities of data, information, knowledge, and wisdom. (b) The relative acquisition cost of the different layers. (c) The gap between data and knowledge and (d) the gap between knowledge and wisdom.

system. In humoral immunity, B cells produce antibodies for neutralization of extracellular pathogens and their antigens that prevent the spread of infection. The activation of B cells and their differentiation into antibody-secreting plasma cells is triggered by antigens and usually requires helper T cells [3]. B cells identify antigens through B-cell receptors, which recognize discrete sites on the surface of target antigens called B-cell epitopes [4].

Cellular immunity involves the activation of phagocytes, antigen-specific cytotoxic T-lymphocytes (CTLs), and the release of various cytokines in response to pathogens and their antigens. T cells identify foreign antigens through their T-cell receptors (TCRs), which interact with a peptide antigen in complex with a major histocompatibility complex (MHC) molecule in conjunction with CD4 or CD8 coreceptors [5, 6]. Peptides that induce immune responses, when presented by MHC on the cell surface for recognition by T cells, are called T-cell epitopes. CD8+ T cells control infection through direct cytolysis of infected cells and through production of soluble antiviral mediators. This function is mediated by linear peptide epitopes presented by MHC class I molecules. CD4+ T cells recognize epitopes presented by MHC class II molecules on the surface of infected cells and secrete lymphokines that stimulate B cells and cytotoxic T cells. The Immune Epitope Database (IEDB) [7] hosts nearly 20,000 T-cell epitopes as of Feb. 2014.

The recognition of a given antigenic peptide by an individual immune system depends on the ability of this peptide to bind one or more of the host's human leukocyte antigens (HLA-human MHC). The binding of antigenic peptides to HLA molecules is the most selective step in identifying T-cell epitopes. There is a great diversity of HLA genes with more than 10,000 known variants characterized as of Feb. 2014 [8]. To manage this diversity, the classification of HLA into supertypes was proposed to describe those HLA variants that have small differences in their peptide-binding grooves and share similar peptide-binding specificities [9, 10]. Peptides that can bind multiple HLA variants are termed "promiscuous peptides." They are suitable for the design of epitope-based vaccines because they can interact with multiple HLA within human populations.

The concept of reverse vaccinology supports identification of vaccine targets by large-scale bioinformatics screening of entire pathogenic genomes followed by experimental validation [11]. Using bioinformatics analysis to select a small set of key wet-lab experiments for vaccine design is becoming a norm. The complexity of identification of broadly protective vaccine targets arises from two principal sources, the diversity of pathogens and the diversity of human immune system. The design of broadly protective peptide-based vaccines involves the identification and selection of vaccine targets composed of conserved T-cell and B-cell epitopes that are broadly cross-reactive to viral subtypes and protective of a large host population (Figure 2).

Fuelled by the breakthroughs in genomics and proteomics and advances in instrumentation, sample processing, and immunological assays, immunology research is entering the big data era. These data vary in granularity, quality, and complexity and are stored in various formats, including publications, technical reports, and databases. Next generation sequencing technologies are shifting the paradigm of genomics and allowing researchers to perform genome-wide studies [12]. It was estimated that the amount of publically available genomic data will grow from petabytes (10^{15}) to exabytes (10^{18}) [13]. Mass spectrometry (MS) is the method for detection and quantitation of proteins. The technical advancements in proteomics support exponential growth of the numbers of characterized protein sequences. It is estimated that more than 2 million protein variants make the posttranslated human proteome in any human individual [14]. Capitalizing on the recent advances in immune profiling methods, the Human Immunology Project Consortium (HIPC) is creating large data sets on human subjects undergoing influenza vaccination or who are infected with pathogens including influenza virus, West Nile virus, herpes zoster, pneumococcus, and the malaria parasite [15]. Systems biology aims to study the interactions between relevant molecular components and their changes over time and enable the development of predictive models. The advent of technological breakthroughs in the fields of genomics, proteomics, and other "omics" is catalyzing advances in systems immunology, a new field under the umbrella of system biology [16]. The synergy between systems immunology and vaccinology enables rational vaccine design [17].

Big data describes the environment where massive data sources combine both structured and unstructured data so that the analysis cannot be performed using traditional database and analytical methods. Increasingly, data sources from literature and online sources are combined with the traditional types of data [18] for summarization of complex information, extraction of knowledge, decision support, and predictive analytics. With the increase of the data sources, both the knowledge and application gaps (Figures 1(c) and 1(d)) keep widening and the corresponding volumes of data and information are rapidly increasing. We describe

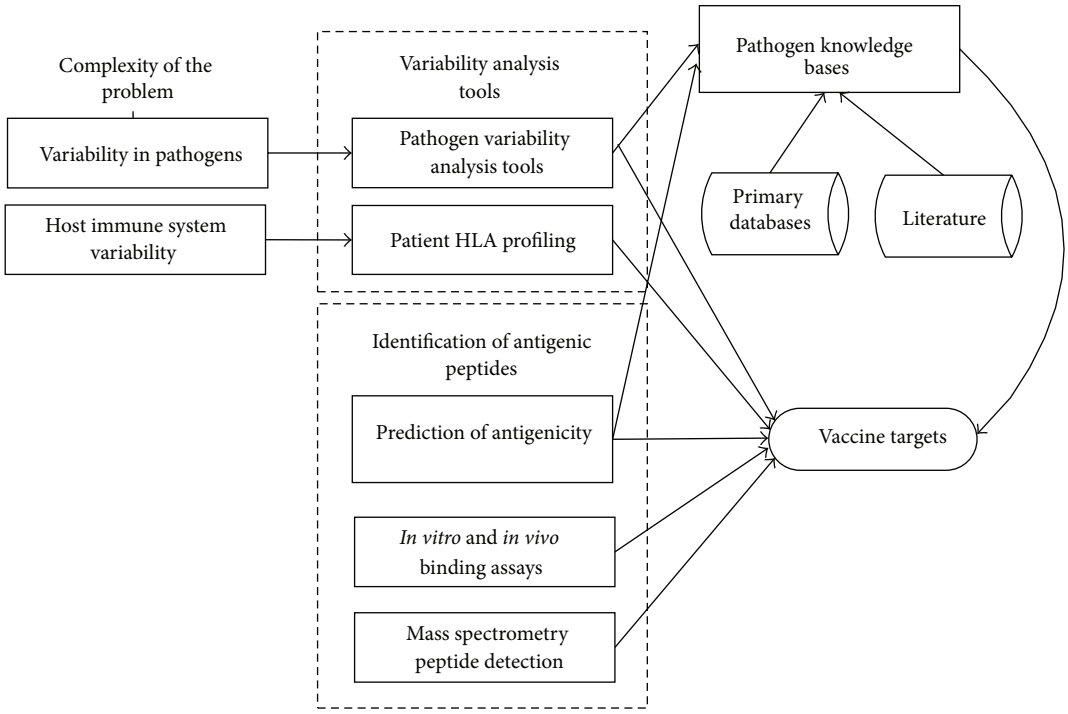


Figure 2: The process of rational vaccine discovery using knowledge-based systems. The design of broadly protective peptide-based vaccines involves identification and selection of vaccine targets composed of conserved T-cell and B-cell epitopes that are broadly cross-reactive to pathogen subtypes and protective of a large host population.

a knowledge-based approach that helps reduce the knowledge and application gaps for applications in immunology and vaccinology.

2. Materials and Methods

In the big data era, knowledge-based systems (KBSs) are emerging as knowledge discovery platforms. A KBS is an intelligent system that employs a computationally tractable knowledgebase or repository in order to reason upon data in a targeted domain and reproduce expert performance relative to such reasoning operations [19]. The goal of a KBS is to increase the reproducibility, scalability, and accessibility of complex reasoning tasks [20]. Some of the web-accessible immunological databases, such as Cancer Immunity Peptide Database that hosts four static data tables containing four types of tumor antigens with defined T-cell epitopes, focus on cataloging the data and information and pay little attention to the integration of analysis tools [21, 22]. Most recent web-accessible immunological databases, such as Immune Epitope Database (IEDB) that catalogs experimentally characterized B-cell and T-cell epitopes and data on MHC binding and MHC ligand elution experiments, started to integrate some data analysis tools [7, 23]. To bridge the knowledge gap between immunological information and knowledge, we need KBSs that tightly integrate data with analysis tools to enable comprehensive screening of immune epitopes from a comprehensive landscape of a given disease (such as influenza, flaviviruses, or cancer), the analysis of crossreactivity and crossprotection following immunization or vaccination, and prediction of neutralizing immune responses. We developed a framework called KB-builder to facilitate data mining by enabling fast development and deployment of web-accessible immunological data knowledge warehouses. The framework consists of seven major functional modules (Figure 3), each facilitating a specific aspect of the knowledgebase construction process. The KB-builder framework is generic and can be applied to a variety of immunological sequence datasets. Its aim is to enable the development of a web-accessible knowledgebase and its corresponding analytics pipeline within a short period of time (typically within 1-2 weeks), given a set of annotated genetic or protein sequences.

The design of a broadly protective peptide-based vaccine against viral pathogens involves the identification and selection of vaccine targets composed of conserved T-cell and B-cell epitopes that are broadly cross-reactive to a wide range of viral subtypes and are protective in a large majority of host population (Figure 2). The KB-builder facilitates a systematic discovery of vaccine targets by enabling fast development of specialized bioinformatics KBS that tightly integrate the content (accurate, up-to-date, and well-organized antigen data) with tailored analysis tools.

The input to KB-builder is data scattered across primary databases and scientific literature (Figure 3). Module 1 (data collection and processing module) performs automated data extraction and initial transformations. The raw antigen

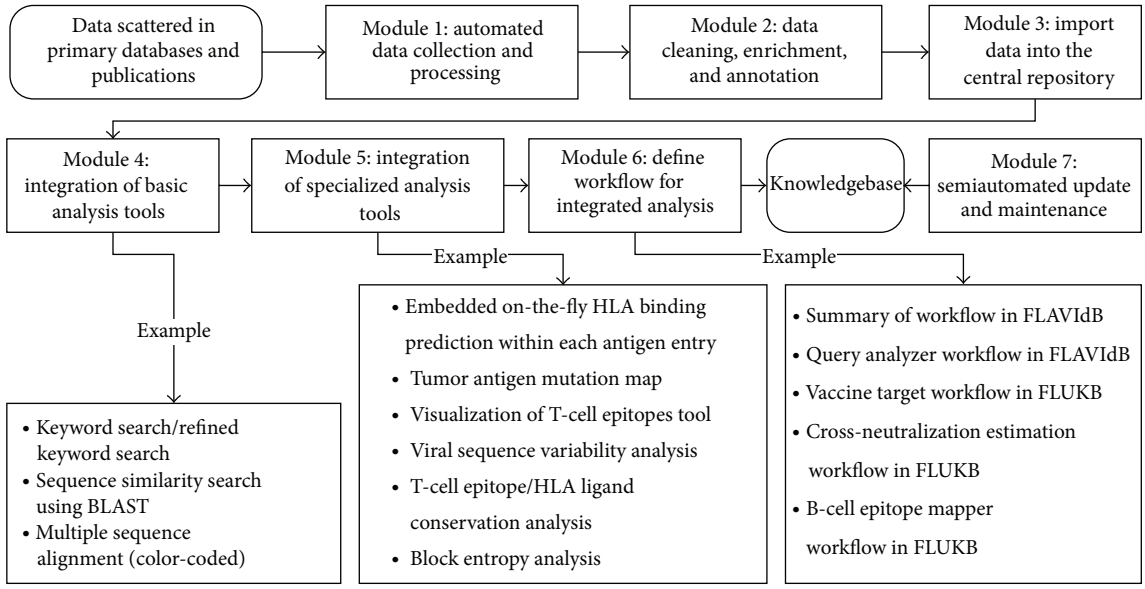


Figure 3: The structure of KB-builder.

data (viral or tumor) consisting of protein or nucleotide sequences, or both, and their related information are collected from various sources. The collected data are then reformatted and organized into a unified XML format. Module 2 (data cleaning, enrichment, and annotation module) deals with data incompleteness, inconsistency, and ambiguities due to the lack of submission standards in the online primary databases. The semiautomated data cleaning is performed by domain experts to ensure data quality, completeness, and redundancy reduction. Semiautomated data enrichment and annotation are performed by the domain experts further enhancing data quality. The semiautomation involves automated comparison of new entries to the entries already processed within the KB and comparison of terms that are entered into locally implemented dictionaries. Terms that match the existing record annotations and dictionary terms are automatically processed. New terms and new annotations are inspected by a curator and if in error they are corrected, or if they represent novel annotations or terms they are added to the knowledgebase and to the local dictionaries. Module 3 (the import module) performs automatic import of the XML file into the central repository. Module 4 (the basic analysis toolset) facilitates fast integration of common analytical tools with the online antigen KB. All our knowledgebases have the basic keyword search tools for locating antigens and T-cell epitopes or HLA ligands. The advanced keyword search tool was included in FLAVIdB, FLUKB, and HPVdB, where users further restrict the search by selecting virus species, viral subtype, pathology, host organism, viral strain type, and several other filters. Other analytical tools include sequence similarity search enabled by basic local alignment search tool (BLAST) [24] and color-coded multiple sequence alignment (MSA) tool [25] on user-defined sequence sets as shown in Figure 4. Module 5 (the specialized analysis toolset) facilitates fast integration of specialized analysis tools designed according to the specific purpose of the knowledgebase and the structural and functional properties of the source of the sequences. To facilitate efficient antigenicity analysis, in every knowledgebase and within each antigen entry, we embedded a tool that performs on-the-fly binding prediction to 15 frequent HLA class I and class II alleles. In TANTIGEN, an interactive visualization tool, mutation map, has been implemented to provide a global view of all mutations reported in a tumor antigen. Figure 5 shows a screenshot of mutation map of tumor antigen epidermal growth factor receptor (EGFR) in TANTIGEN. In TANTIGEN and HPVdB, a T-cell epitope visualization tool has been implemented to display epitopes in all isoforms of a tumor antigen or sequences of a HPV genotype. The B-cell visualization tool in FLAVIdB and FLUKB displays neutralizing B-cell epitope positions on viral protein three-dimensional (3D) structures [26, 27]. To analyze viral sequence variability, given a MSA of a set of sequences, a tool was developed to calculate Shannon entropy at each alignment position. To identify conserved T-cell epitopes that cover the majority of viral population, we developed and integrated block entropy analysis tool in FLAVIdB and FLUKB to analyze peptide conservation and variability. We developed a novel sequence logo tool, BlockLogo, optimized for visualization of continuous and discontinuous motifs, fragments [28, 29]. When paired with the HLA binding prediction tool, BlockLogo is a useful tool for rapid assessing of immunological potential of selected regions in a MSA, such as alignments of viral sequences or tumor antigens.

A workflow is an automated process that takes a request from the user, performs complex analysis by combining data and tools preselected for common questions, and produces a comprehensive report [30]. Module 6 (workflow for integrated analysis to answer meaningful questions) automates the consecutive execution of multiple analysis steps, which researchers usually would have to perform manually, to answer complex sequential questions. Two workflow types,

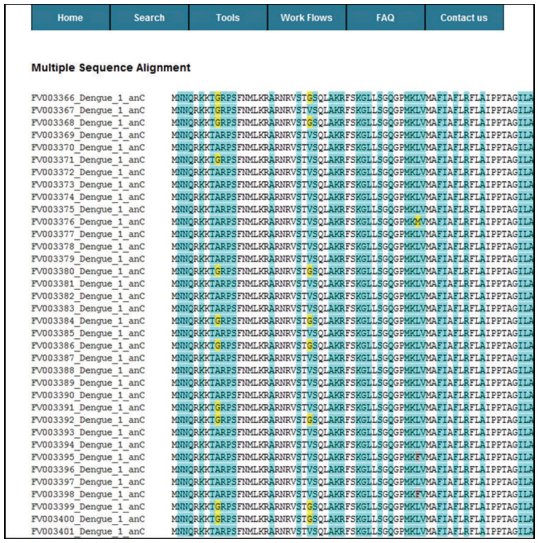


Figure 4: A screenshot of the result page generated by the color-coded MSA tool implemented in the FLAVIdB. The residues are color-coded by frequency: white (100%), cyan (second most frequent), yellow (third most frequent residues), gray (fourth most frequent residues), green (fifth most frequent residues), purple (sixth most frequent residues), and blue (everything less frequent than the sixth most frequent residues).

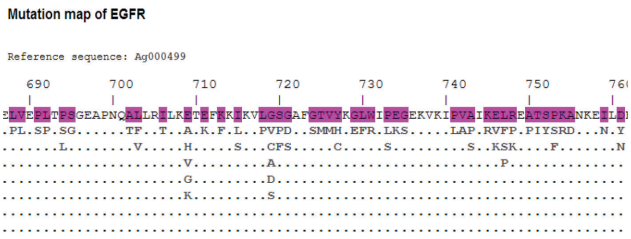


Figure 5: A screenshot of mutation map of tumor antigen epidermal growth factor receptor (EGFR) in TANTIGEN. The numbers are the amino acid positions in the antigen sequence and the top amino acid sequence is the reference sequence of EGFR. The highlighted amino acids in the reference sequences are positions where point mutations took place. Clicking on the amino acids below the point mutation positions links to the mutated sequence data table.

the summary workflow and the query analyzer workflow, were implemented in FLAVIdB. Three workflow types, the vaccine target workflow, the crossneutralization estimation workflow, and B-cell epitope mapper workflow, were implemented in FLUKB. Module 7 (semiautomated update and maintenance of the databases) employs a semiautomated approach to maintain and update the databases.

3. Results and Discussion

Using the KB-builder, we built several immunovaccinology knowledgebases including TANTIGEN: Tumor T-cell Antigen Database (http://cvc.dfci.harvard.edu/tadb/), FLAVIdB: Flavivirus Antigen Database [31], HPVdB: Human Papillomavirus T-cell Antigen Database [32], FLUKB: Flu

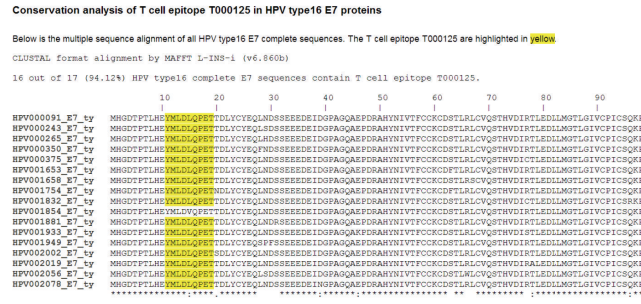
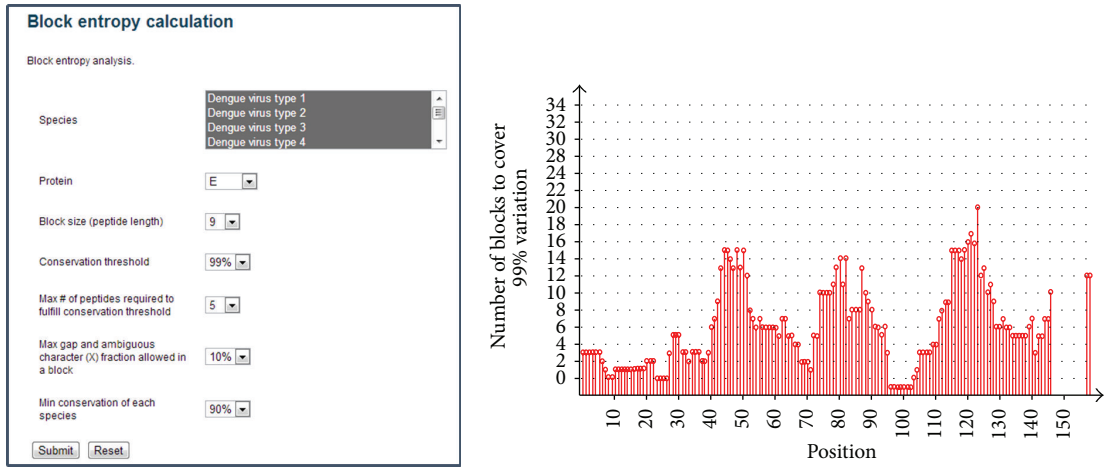


Figure 6: A screenshot of the conservation analysis result page of T-cell epitope $E7_{11-19}$ in HPVdB.

Virus Antigen Database (http://research4.dfci.harvard.edu/cvc/flukb/), Epstein-Barr Virus T-cell Antigen Database (http://research4.dfci.harvard.edu/cvc/ebv/), and Merkel Cell Polyomavirus Antigen Database (http://cvc.dfci.harvard.edu/mcv/). These knowledgebases combine virus and tumor antigenic data, specialized analysis tools, and workflow for automated complex analyses focusing on applications in immunology and vaccinology.

The Human Papillomavirus T-cell Antigen Database (HPVdB) contains 2781 curated antigen entries of antigenic proteins derived from 18 genotypes of high-risk HPV and 18 genotypes of low-risk HPV. It also catalogs 191 verified T-cell epitopes and 45 verified HLA ligands. The functions of the data mining tools integrated in HPVdB include antigen and epitope/ligand search, sequence comparison using BLAST search, multiple alignments of antigens, classification of HPV types based on cancer risk, T-cell epitope prediction, T-cell epitope/HLA ligand visualization, T-cell epitope/HLA ligand conservation analysis, and sequence variability analysis.

HPV regulatory proteins E6 and E7 proteins are often studied for immune-based therapies as they are constitutively expressed in HPV-associated cancer cells. First, the prediction of A∗0201 binding peptides (both 9-mers and 10-mers) of HPV16 E6 and E7 proteins was performed computationally. Based on the prediction results, 21 peptides were synthesized and ten of them were identified as binders using an A∗0201 binding assay. The ten A∗0201-binding peptides were further tested for immune recognition in peripheral blood mononuclear cells isolated from six A∗0201-positive healthy donors using interferon γ (IFN γ) ELISpot assay. Two peptides, $E7_{11-19}$ and $E6_{29-38}$, elicited spot-forming-unit numbers 4-5-fold over background in one donor. Finally, mass spectrometry was used to validate that peptide $E7_{11-19}$ is naturally presented on HPV16-transformed, A∗0201-positive cells. Using the peptide conservation analysis tool embedded in HPVdB, we answered the question how many HPV strains contain this epitope. The epitope $E7_{11-19}$ is conserved in 16 of 17 (94.12% conserved) HPV16 E7 complete sequences (Figure 6). A single substitution mutation L15V in HPV001854 (UniProt ID: C0KXQ5) resulted in the immune escape. Among the 35 HPV16 cervical cancer samples we analyzed, only a single sample contained the HPV001854 sequence variant. The conserved HPV T-cell epitopes displayed by HPV transformed tumors such as $E7_{11-19}$ may be the basis of



(a) (b)

Figure 7: Block entropy analysis of envelope proteins of dengue subtypes 1–4 in the FLAVIdB. (a) A screenshot of the input page of block entropy analysis in the FLAVIdB. (b) The number of blocks needed to cover 99% of the sequences variation. x-axis is the starting positions of blocks and y-axis is the number of blocks required. The blocks with gap fraction above 10% are not plotted.

a therapeutic T-cell based cancer vaccine. This example shows the combination of bioinformatics analysis and experimental validation leading to identification of suitable vaccine targets [33, 34].

Flaviviruses, such as dengue and West Nile viruses, are NIAID Category A and B Priority Pathogens. We developed FLAVIdB that contains 12,858 entries of flavivirus antigen sequences, 184 verified T-cell epitopes, 201 verified B-cell epitopes, and 4 representative molecular structures of the dengue virus envelope protein [31]. The data mining system integrated in FLAVIdB includes tools for antigen and epitope/ligand search, sequence comparison using BLAST search, multiple alignments of antigens, variability and conservation analysis, T-cell epitope prediction, and characterization of neutralizing components of B-cell epitopes. A workflow is an automated process that takes a request from the user, performs complex analysis by combining data and tools preselected for common questions, and produces a comprehensive report to answer a specific research question. Two predefined analysis workflow types, summary workflow and query analyzer workflow, were implemented in FLAVIdB [31].

Broad coverage of the pathogen population is particularly important when designing T-cell epitope vaccines against viral pathogens. Using FLAVIdB we applied the block entropy analysis method to the proteomes of the four serotypes of dengue virus (DENV) and found 1,551 blocks of 9-mer peptides, which cover 99% of available sequences with five or fewer unique peptides [35]. Many of the blocks are located consecutively in the proteins, so connecting these blocks resulted in 78 conserved regions which can be covered with 457 subunit peptides. Of the 1551 blocks of 9-mer peptides, 110 blocks consisted of peptides all predicted to bind to MHC with similar affinity and the same HLA restriction. In total, we identified a pool of 333 peptides as T-cell epitope candidates. This set could form the basis for a broadly neutralizing dengue virus vaccine. The results of block entropy analysis of dengue subtypes 1–4 from FLAVIdB are shown in Figure 7.

Influenza virus is a NIAID Category C Priority Pathogen. We developed the FLUKB that currently contains 302,272 influenza viral protein sequence entries from 62,016 unique strains (57,274 type A, 4,470 type B, 180 type C, and 92 unknown types) of influenza virus. It also catalogued 349 unique T-cell epitopes, 708 unique MHC binding peptides, and 17 neutralizing antibodies against hemagglutinin (HA) proteins along with their 3D structures. The detailed information on the neutralizing antibodies such as isolation information, experimentally validated neutralizing/escape influenza strains, B-cell epitope on the 3D structures, are also provided.

Approximately 10% of B-cell epitopes are linear peptides, while 90% are formed from discontinuous amino acids that create surface patches resulting from 3D folding of proteins [36]. Characterization of an increasing number of broadly neutralizing antibodies specific for pathogen surface proteins, the growing number of known 3D structures of antigen-neutralizing antibody complexes, and the rapid growth of the number of viral variant sequences demand systematic bioinformatics analyses of B-cell epitopes and cross-reactivity of neutralizing antibodies. We developed a generic method for the assessment of neutralizing properties of monoclonal antibodies. Previously, dengue virus was used to demonstrate a generalized method [27]. This methodology has direct relevance to the characterization and the design of broadly neutralizing vaccines.

Using the FLUKB, we employed the analytical methods to estimate cross-reactivity of neutralizing antibodies (nAbs) against surface glycoprotein HA of influenza virus strains, both newly emerging or the existing ones [26]. We developed a novel way of describing discontinuous motifs as virtual peptides to represent B-cell epitopes and to estimate potential

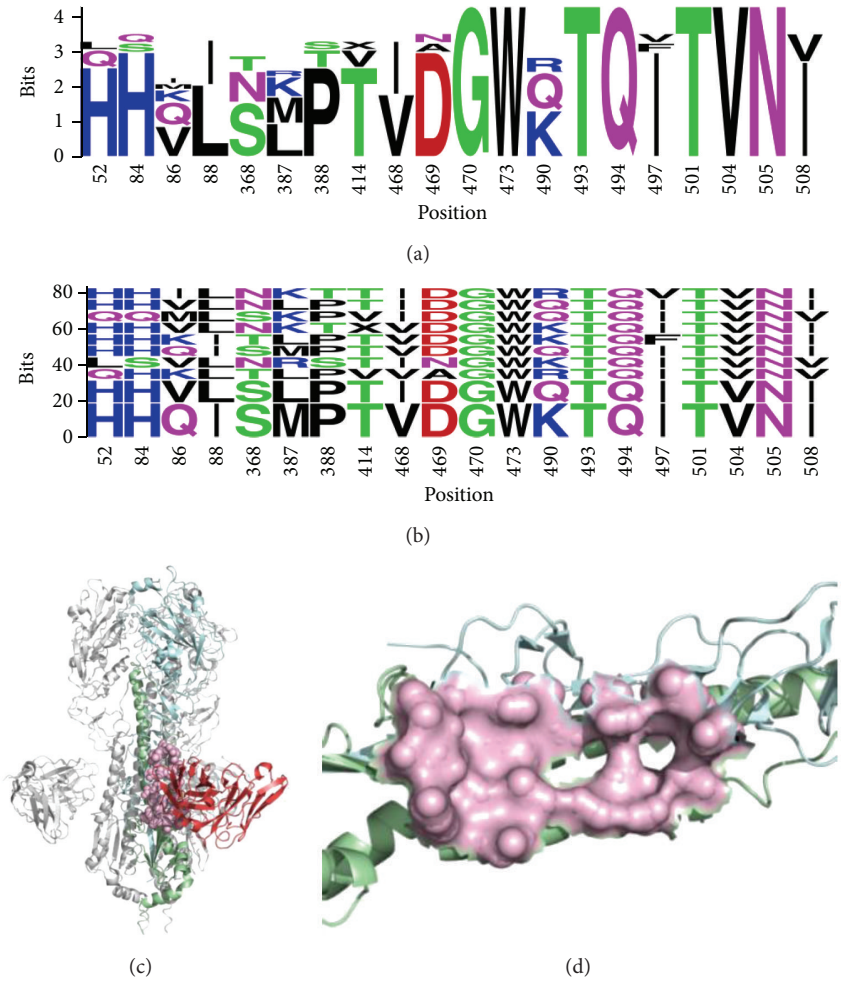


Figure 8: (a) Sequence logo of neutralizing epitopes by neutralizing antibody F10 on influenza virus HA protein. (b) BlockLogo of the discontinuous residues in F10 neutralizing epitope. (c) The structure of influenza A HA protein with neutralizing antibody F10 (PDB ID:3FKU) and the conformational epitope shown in pink. (d) Discontinuous epitope on HA protein recognized by F10.

cross-reactivity and neutralizing coverage of these epitopes. Strains labelled as potentially cross-reactive are those that share 100% identity of B-cell epitopes with experimentally verified neutralized strains. Two workflow types were implemented in the FLUKB for cross-neutralization analysis: cross-neutralization estimation workflow and B-cell epitope mapper workflow.

The cross-neutralization estimation workflow estimates the cross-neutralization coverage of a validated neutralizing antibody using all full-length sequences of HA hosted in the FLUKB, or using full-length HA sequences of a user-defined subset by restricting year ranges, subtypes, or geographical locations. Firstly, a MSA is generated using the full-length HA sequences. The resulting MSA provides a consistent alignment position numbering scheme for the downstream analyses. Secondly, for each nAb, the HA sequence from its 3D structure and from the experimentally validated strains is used to search for a strain with the highest similarity in FLUKB using BLAST. Thirdly, a B-cell epitope is identified from the validated antigen-antibody structures based on the calculation of accessible surface area and atom distance. Fourthly, using the MSA and the alignment position numbering, the residue position of the B-cell epitope is mapped onto the HA sequences of validated strains to get B-cell epitope motifs. Discontinuous motifs are extracted from all the HA sequences in the MSA and compared to the B-cell epitope motif. According to the comparison results, they are classified to be either neutralizing if identical to a neutralizing discontinuous motif, escape if identical to an escape discontinuous motif, or not validated if no identical match was found. The cross-neutralization coverage estimation of neutralizing antibody F10 on all HA sequences from FLUKB is shown in Figure 8.

For a newly emerged strain, the B-cell epitope mapper workflow performs *in silico* prediction of its cross-neutralization based on existing nAbs and provides preliminary results for the design of downstream validation experiments. Firstly, a discontinuous peptide is extracted from its HA sequence according to positions on each known B-cell epitope. Secondly, sequence similarity comparison

is conducted between the discontinuous motifs and all known B-cell epitopes from experimentally validated strains. The motifs identical to the known neutralized or escape B-cell epitope motifs are proposed as neutralized or escape strains, respectively.

The cross-neutralization estimation workflow provides an overview of cross-neutralization of existing neutralizing antibodies, while B-cell epitope mapper workflow gives an estimation of possible neutralizing effect of new viral strains using known neutralizing antibodies. This knowledge-based approach improves our understanding of antibody/antigen interactions, facilitates mapping of the known universe of target antigens, allows the prediction of cross-reactivity, and speeds up the design of broadly protective influenza vaccines.

4. Conclusions

The big data analytics applies advanced analytic methods to data sets that are very large and complex and that include diverse data types. These advanced analytics methods include predictive analytics, data mining, text mining, integrated statistics, visualization, and summarization tools. The data sets used in our case studies are complex and the analytics is achieved through the definition of workflow. Data explosion in our case studies is fueled by the combinatorial complexity of the domain and the disparate data types. The cost of analysis and computation increases exponentially as we combine various types of data to answer research questions. We use the *in silico* identification of influenza T-cell epitopes restricted by HLA class I variants as an example. There are 300,000 influenza sequences to be analyzed for T-cell epitopes using MHC binding prediction tools based on artificial neural networks or support vector machines [37–40]. Based on the DNA typing for the entire US donor registry, there are 733 HLA-A, 921 HLA-B, and 429 HLA-C variants, a total of 2083 HLA variants, observed in US population [41]. These alleles combine into more than 45,000 haplotypes (combinations of HLA-A, -B, and -C) [41]. Each of these haplotypes has different frequencies and distributions across different populations. The *in silico* analysis of MHC class I restricted T-cell epitopes includes MHC binding prediction of all overlapping peptides that are 9–11 amino acids long. This task alone involves a systematic analysis of 300,000 sequences that are on average 300 amino acids long. Therefore, the total number of *in silico* predictions is approximately $300{,}000 \times 300 \times 3 \times 2083$ (number of sequences times the average length of each sequence times 3 times the number of observed HLA variants) or a total of 5.6×10^{11} calculations. Predictive models do not exist for all HLA alleles, so some analysis needs to be performed by analysis of similarity of HLA molecules and grouping them in clusters that share binding properties. For B-cell epitope analysis, the situation is similar, except that the methods involve the analysis of 3D structures of antibodies and the analysis of nearly 100,000 sequences of HA and neuraminidase (NA) and their cross-comparison for each neutralizing antibody. A rich set of visualization tools is needed to report population data and distributions across populations. For vaccine studies, these data need to be analyzed together with epidemiological data including transmissibility and severity of influenza viruses [42]. These functional properties can be assigned to each influenza strain and the analysis can be performed for their epidemic and pandemic potential. These numbers indicate that the analytics methods involve a large amount of calculations that cannot be performed using brute force approaches.

Immunological knowledge discovery relies heavily on both the availability of accurate, up-to-date, and well-organized data and the proper analytics tools. We propose the use of knowledge-based approaches by developing knowledgebases combining well-annotated data with specialized analytical tools and integrating them into analytical workflow. A set of well-defined workflow types with rich summarization and visualization capacity facilitates the transformation from data to critical information and knowledge. By using KB-builder, we enabled streamlining of normally time-consuming process of database development. The knowledgebases built using KB-builder will speed up rational vaccine design by providing accurate and well-annotated data coupled with tailored computational analysis tools and workflow.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. Rowley, "The wisdom hierarchy: representations of the DIKW hierarchy," *Journal of Information Science*, vol. 33, no. 2, pp. 163–180, 2007.

[2] R. Ackoff, "From data to wisdom," *Journal of Applies Systems Analysis*, vol. 16, no. 1, pp. 3–9, 1989.

[3] C. Janeway, *Immunobiology: The Immune System in Health and Disease*, Garland Science, New York, NY, USA, 6th edition, 2005.

[4] M. H. V. van Regenmortel, "What is a B-cell epitope?" *Methods in Molecular Biology*, vol. 524, pp. 3–20, 2009.

[5] S. C. Meuer, S. F. Schlossman, and E. L. Reinherz, "Clonal analysis of human cytotoxic T lymphocytes: T4+ and T8+ effector T cells recognize products of different major histocompatibility complex regions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 79, no. 14 I, pp. 4395–4399, 1982.

[6] J. H. Wang and E. L. Reinherz, "Structural basis of T cell recognition of peptides bound to MHC molecules," *Molecular Immunology*, vol. 38, no. 14, pp. 1039–1049, 2002.

[7] R. Vita, L. Zarebski, J. A. Greenbaum et al., "The immune epitope database 2.0," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D854–D862, 2009.

[8] J. Robinson, J. A. Halliwell, H. McWilliam, R. Lopez, P. Parham, and S. G. E. Marsh, "The IMGT/HLA database," *Nucleic Acids Research*, vol. 41, no. 1, pp. D1222–D1227, 2013.

[9] A. Sette and J. Sidney, "Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism," *Immunogenetics*, vol. 50, no. 3-4, pp. 201–212, 1999.

[10] O. Lund, M. Nielsen, C. Kesmir et al., "Definition of supertypes for HLA molecules using clustering of specificity matrices," *Immunogenetics*, vol. 55, no. 12, pp. 797–810, 2004.

[11] R. Rappuoli, "Reverse vaccinology," *Current Opinion in Microbiology*, vol. 3, no. 5, pp. 445–450, 2000.

[12] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, "The next-generation sequencing revolution and its impact on genomics," *Cell*, vol. 155, no. 1, pp. 27–38, 2013.

[13] D. R. Zerbino, B. Paten, and D. Haussler, "Integrating genomes," *Science*, vol. 336, no. 6078, pp. 179–182, 2012.

[14] M. Uhlen and F. Ponten, "Antibody-based proteomics for human tissue profiling," *Molecular and Cellular Proteomics*, vol. 4, no. 4, pp. 384–393, 2005.

[15] V. Brusic, R. Gottardo, S. H. Kleinstein, and M. M. Davis, "Computational resources for high-dimensional immune analysis from the human immunology project consortium," *Nature Biotechnology*, vol. 32, no. 2, pp. 146–148, 2014.

[16] A. Aderem, "Editorial overview: system immunology," *Seminars in Immunology*, vol. 25, no. 3, pp. 191–192, 2013.

[17] S. Li, H. I. Nakaya, D. A. Kazmin, J. Z. Oh, and B. Pulendran, "Systems biological approaches to measure and understand vaccine immunity in humans," *Seminars in Immunology*, vol. 25, no. 3, pp. 209–218, 2013.

[18] L. Olsen, U. J. Kudahl, O. Winther, and V. Brusic, "Literature classification for semi-automated updating of biological knowledgebases," *BMC Genomics*, vol. 14, supplement 5, article S14, 2013.

[19] P. R. O. Payne, "Chapter 1: biomedical knowledge integration," *PLoS Computational Biology*, vol. 8, no. 12, Article ID e1002826, 2012.

[20] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications—a decade review from 2000 to 2011," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11303–11311, 2012.

[21] N. Vigneron, V. Stroobant, B. J. van den Eynde, and P. van der Bruggen, "Database of T cell-defined human tumor antigens: the 2013 update," *Cancer Immunity*, vol. 13, article 15, 2013.

[22] B. J. van den Eynde and P. van der Bruggen, "T cell defined tumor antigens," *Current Opinion in Immunology*, vol. 9, no. 5, pp. 684–693, 1997.

[23] B. Peters, J. Sidney, P. Bourne et al., "The design and implementation of the immune epitope database and analysis resource," *Immunogenetics*, vol. 57, no. 5, pp. 326–336, 2005.

[24] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[25] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[26] J. Sun, U. J. Kudahl, C. Simon, Z. Cao, E. L. Reinherz, and V. Brusic, "Large-scale analysis of B-cell epitopes on influenza virus hemagglutinin—implications for cross-reactivity of neutralizing antibodies," *Frontiers in Immunology*, vol. 5, article 38, 2014.

[27] J. Sun, G. L. Zhang, L. R. Olsen, E. L. Reinherz, and V. Brusic, "Landscape of neutralizing assessment of monoclonal antibodies against dengue virus," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics (BCB '13)*, p. 836, Washington, DC, USA, 2013.

[28] G. E. Crooks, G. Hon, J. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Research*, vol. 14, no. 6, pp. 1188–1190, 2004.

[29] L. R. Olsen, U. J. Kudahl, C. Simon et al., "BlockLogo: visualization of peptide and sequence motif conservation," *Journal of Immunological Methods*, vol. 400-401, pp. 37–44, 2013.

[30] J. Söllner, A. Heinzel, G. Summer et al., "Concept and application of a computational vaccinology workflow," *Immunome Research*, vol. 6, supplement 2, article S7, 2010.

[31] L. R. Olsen, G. L. Zhang, E. L. Reinherz, and V. Brusic, "FLAVIdB: a data mining system for knowledge discovery in flaviviruses with direct applications in immunology and vaccinology," *Immunome Research*, vol. 7, no. 3, pp. 1–9, 2011.

[32] G. L. Zhang, A. B. Riemer, D. B. Keskin, L. Chitkushev, E. L. Reinherz, and V. Brusic, "HPVdb: a data mining system for knowledge discovery in human papillomavirus with applications in T cell immunology and vaccinology," *Database*, vol. 2014, Article ID bau031, 2014.

[33] A. B. Riemer, D. B. Keskin, G. Zhang et al., "A conserved E7-derived cytotoxic T lymphocyte epitope expressed on human papillomavirus 16-transformed HLA-A2+ epithelial cancers," *Journal of Biological Chemistry*, vol. 285, no. 38, pp. 29608–29622, 2010.

[34] D. B. Keskin, B. Reinhold, S. Lee et al., "Direct identification of an HPV-16 tumor antigen from cervical cancer biopsy specimens," *Frontiers in Immunology*, vol. 2, article 75, 2011.

[35] L. R. Olsen, G. L. Zhang, D. B. Keskin, E. L. Reinherz, and V. Brusic, "Conservation analysis of dengue virust-cell epitope-based vaccine candidates using peptide block entropy," *Frontiers in Immunology*, vol. 2, article 69, 2011.

[36] J. Huang and W. Honda, "CED: a conformational epitope database," *BMC Immunology*, vol. 7, article 7, 2006.

[37] E. Karosiene, M. Rasmussen, T. Blicher, O. Lund, S. Buus, and M. Nielsen, "NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ," *Immunogenetics*, vol. 65, no. 10, pp. 711–724, 2013.

[38] I. Hoof, B. Peters, J. Sidney et al., "NetMHCpan, a method for MHC class i binding prediction beyond humans," *Immunogenetics*, vol. 61, no. 1, pp. 1–13, 2009.

[39] G. L. Zhang, I. Bozic, C. K. Kwoh, J. T. August, and V. Brusic, "Prediction of supertype-specific HLA class I binding peptides using support vector machines," *Journal of Immunological Methods*, vol. 320, no. 1-2, pp. 143–154, 2007.

[40] G. L. Zhang, A. M. Khan, K. N. Srinivasan, J. T. August, and V. Brusic, "Neural models for predicting viral vaccine targets," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 5, pp. 1207–1225, 2005.

[41] L. Gragert, A. Madbouly, J. Freeman, and M. Maiers, "Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry," *Human Immunology*, vol. 74, no. 10, pp. 1313–1320, 2013.

[42] C. Reed, M. Biggerstaff, L. Finelli et al., "Novel framework for assessing epidemiologic effects of influenza epidemics and pandemics," *Emerging Infectious Diseases*, vol. 19, no. 1, pp. 85–91, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 943186, 8 pages
http://dx.doi.org/10.1155/2014/943186

Research Article

Geometric Analysis of Alloreactive HLA α-Helices

Reiner Ribarics,[1] Rudolf Karch,[1] Nevena Ilieva,[2] and Wolfgang Schreiner[1]

[1] *Section of Biosimulation and Bioinformatics, Center for Medical Statistics, Informatics and Intelligent Systems (CeMSIIS), Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria*
[2] *Institute for Nuclear Research and Nuclear Energy (INRNE), Bulgarian Academy of Sciences 72, Tzarigradsko Chaussee, Sofia 1784, Bulgaria*

Correspondence should be addressed to Wolfgang Schreiner; wolfgang.schreiner@meduniwien.ac.at

Received 31 January 2014; Accepted 21 May 2014; Published 17 June 2014

Academic Editor: Francesco Pappalardo

Molecular dynamics (MD) is a valuable tool for the investigation of functional elements in biomolecules, providing information on dynamic properties and processes. Previous work by our group has characterized static geometric properties of the two MHC α-helices comprising the peptide binding region recognized by T cells. We build upon this work and used several spline models to approximate the overall shape of MHC α-helices. We applied this technique to a series of MD simulations of alloreactive MHC molecules that allowed us to capture the dynamics of MHC α-helices' steric configurations. Here, we discuss the variability of spline models underlying the geometric analysis with varying polynomial degrees of the splines.

1. Introduction

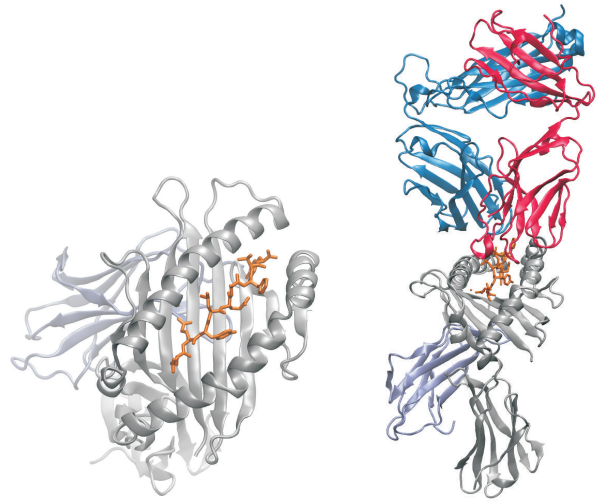
Major histocompatibility complexes (MHC) play a key role in immune reactions. The function of this class of highly polymorphic proteins is to bind peptide fragments (p) derived from pathogens or tumour antigens and display them on the cell surface for recognition by appropriate T cells. T cells can detect these peptide fragments from pathogens or cancer cells by T cell receptor (TCR) molecules on their cell surface, but only if these peptides are presented in complex with MHC molecules (pMHC). As a consequence of the TCR/pMHC interaction, pathogen-infected cells or cancer cells can be detected and eliminated by the immune system.

The peptide binding region of class I MHC molecules comprises two α-helices and a β-sheet as a floor. The α-helices are orientated in an antiparallel manner to form a binding pocket (see Figure 1).

The TCR is a heterodimer comprising one α- and one β-chain. Each chain has a constant and a variable domain. The constant domain is facing the cell membrane, while the variable domain is facing the extracellular space ready for interaction with MHC molecules. The area of interaction between these two proteins comprises the two MHC α-helices and the three hypervariable complementarity determining regions (CDR) 1, 2, and 3 of the TCR. The peptide in the MHC binding groove mainly interacts with the TCR via CDR 3.

TCR and MHC molecules show wide diversity, therefore sophisticated selection mechanisms exist to prevent autoreactivity that could lead to autoimmunity. During development, T cells are restricted to only recognize host MHC [1, 2]. In other words, T cells only recognize cognate antigen presented by one of the MHC molecules that are present in the host in which they have developed. These T cells, however, may directly react with MHC molecules that are not present in the host (allogeneic reaction). T cells form the basis of allograft rejection, where the host immune system recognizes the transplant as an intruder due to allogeneic MHC molecules.

With molecular dynamics (MD) simulations it is possible to simulate the physical movements of atoms and molecules by solving Newton's equations of motion. The simulations can be used to investigate functional molecular elements and dynamic molecular processes, for example, signal transduction [3–7]. We analyse molecular dynamics simulations of three closely related MHC molecules of the HLA class B44: HLA-B*44:02, HLA-B*44:03, and HLA-B*44:05. Each MHC molecule harbours the self-antigen ABCD3 in its antigen binding groove and is ligated to the LC13 T cell receptor. Of note, the LC13 TCR alloreacts with HLA-B*44:02



(a) (b)

Figure 1: *Molecular structure of MHC class I*. Three-dimensional representation of secondary structural elements of. (a) HLA-B*44:05 (grey), ABCD3 peptide (orange), and β_2-microglobulin (ice blue). (b) HLA-B*44:05 (grey), ABCD3 peptide (orange), β_2-microglobulin (ice blue), and LC13 T cell receptor (dark blue and red). PDB ID: 3KPS.

and HLA-B*44:05, but not with HLA-B*44:03. This fact is also reflected in the binding affinities of the respective TCR/pMHC complexes: LC13 binds HLA-B*44:02 and HLA-B*44:05 with high affinity, whereas binding HLA-B*44:03 is very weak [8].

Previous work by our group characterized the geometric properties of MHC α-helices of a plethora of static crystal structures [9–11] found in the protein database (PDB, [12]). The aim of the present work is to describe the dynamics of the MHC α-helices in the above-mentioned set of allogeneic HLA-types using spline representation. Spline representation allows to mathematically represent the overall shape of the MHC α-helices and capture their geometric properties over the simulation time. The mathematical description of structural elements of macromolecules has been used before, for example, for visualization [13], for calculation of differential geometric parameters of helix bundles [14], and for monitoring of structural changes of Leucine-Rich Repeat (LRR) proteins [15].

We focus on the analysis of geometric mathematical quantities of the MHC α-helices that allow us to characterize their shape and geometry, that is, interhelical distance and area of the ruled surface spanned by the MHC α-helices. We also discuss the variability of spline models with polynomial degrees $m = 2, 3, 4$ and $K = 0$ interior knots.

2. Methods

2.1. Construction of Complexes for Molecular Dynamics Simulation. Conformational transitions occur on a variety of time scales ranging from nanoseconds to seconds [13]. This work represents a proof of concept study for geometrical representation of MHC α-helices, hence, molecular dynamics simulations were performed for 250 ns. A simulation time of 250 ns is a feasible choice for such large systems of about 400.000 atoms (proteins and solvent) as shown in Table 1.

Table 1: Molecular systems simulated.

Molecular system	Simulation length
LC13 TCR/ABCD3/HLA-B*44:02 (B4402)	250 ns
LC13 TCR/ABCD3/HLA-B*44:03 (B4403)	250 ns
LC13 TCR/ABCD3/HLA-B*44:05 (B4405)	250 ns

HLA-B*44:05 plus ABCD3 peptide (EEYLQAFTY) ligated to the LC13 TCR have been successfully crystallized by Macdonald et al. [8]. The structure is accessible on http://www.pdb.org/ (PDB ID 3KPS). Unfortunately, the structures of HLA-B*44:02 and HLA-B*44:03 plus ABCD3 peptide and LC13 TCR have not been resolved. Therefore, we used the technique of homology modelling to create the missing structures.

For generation of LC13/ABCD3/HLA-B*44:03 (complex of TCR/pMHC) we used PDB structure 3KPS as a template. As mentioned above, this structure file includes LC13/ABCD3/HLA-B*44:05. In order to change the HLA type from B*44:05 to B*44:03 we introduced Y116D and D156L mutations into the MHC molecule (amino acid positions specified by PDB numbering; see Figure 2). To modify the crystal structure and substitute amino acids, that is, in silico mutagenesis, we used the Swiss PDB Viewer. This program allows users to change amino acid side-chains and automatically browses a rotamer library to select that rotamer minimising a scoring function. Rotamers are defined as low-energy side-chain conformations. However, the rotamer energy optimisation by the scoring function only works locally and can, in certain circumstances, result in clashes, that is, atoms come into close contact so that the repulsion

HLA numbering	140	160	180
PDB numbering	120	140	160
B44:05/130–200	DGRLLRGYDQ**Y**AYDGKDYIALNEDLSSWTAADTAAQITQRKWEAARVAEQDRAYLEGLCVESLRRYLENGK		
B44:03/130–200	DGRLLRGYDQDAYDGKDYIALNEDLSSWTAADTAAQITQRKWEAARVAEQ**L**RAYLEGLCVESLRRYLENGK		
B44:02/130–200	DGRLLRGYDQDAYDGKDYIALNEDLSSWTAADTAAQITQRKWEAARVAEQDRAYLEGLCVESLRRYLENGK		

Figure 2: Alignment of amino acid sequences of HLA-B*44:02, HLA-B*44:03, and HLA-B*44:05 (downloaded from IMGT/HLA database [14]). HLA-B*44:05 was used as a template, because a three-dimensional structure of this MHC in complex with ABCD3 peptide and LC13 TCR was available. Sequence alignment was done with CLC bio's sequence viewer. Note that PDB sequence numbering and IMGT/HLA database numbering differ.

term of the Lennard-Jones potential predominates. Proper energy minimisation is routinely performed in the subsequent molecular dynamics simulation protocol (we used a steepest-descent method).

For the generation of LC13/ABCD3/HLA-B*44:02 we again used PDB structure 3KPS as a template. In order to change the HLA allele from B*44:05 to B*44:02, we introduced the Y116D mutation into the MHC molecule (see Figure 2 for sequence alignment) using in silico mutagenesis.

2.2. Molecular Dynamics Simulation Protocol. MD simulation of TCR/pMHC systems (B4402, B4403, and B4405) was performed using GROMACS 4.0.7 [15] according to the following protocol.

First, we immersed the TCR/pMHC complex in a SPC [16] artificial water bath (cubic box) allowing for a minimum distance of 2 nm between complex and box boundaries. Second, we added sodium and chloride ions to a concentration of 0.15 mol/L and at the same time neutralized the net charge of the system. Third, we minimized the energy of the solvated system using a steepest descent method. Next, we warmed up the system to 310 K during a 100 ps position-restraints MD simulation. Finally, we carried out MD production runs with LINCS constraint algorithm acting on all bonds and using the GROMOS96 53a6 force field [17]. Hydrogen motions were removed allowing for an integration step of 5 fs. Coordinates were written to disk every 50 ps of simulation time. Coulomb interactions were computed using Particle Mesh Ewald (PME) with a maximum grid spacing of 0.12 nm and interpolation order 4. Van der Waals and Coulomb interactions were computed with a cut-off at 1.4 nm. Velocity rescale temperature coupling was set to 310 K and Berendsen isotropic pressure coupling was set to 1 bar. All other parameters were set in accordance with Omasits et al. [18].

2.3. Spline Representation of MHC α-Helices. The MH^2C software package introduced by Hischenhuber et al. [11] provides a general approach to model α-helices of any macromolecule containing this secondary structural element. Molecular dynamics simulations of TCR/pMHC complexes yield a series of time evolving molecular conformational structures. The resulting structures were subjected to analysis by MH^2C. As mentioned in the introduction, MHC molecules comprise two α-helices, hereafter named G-ALPHA1 helix and G-ALPHA2 helix. In order to mathematically describe and quantify the helical movements, spline curves $\overrightarrow{c}(z)$ are fitted to the α-helices where z is the curve parameter. To do that, we extracted the C_α atom coordinates of the α-helices' amino acids, which are in accordance with the classification of α-helices of visual molecular dynamics (VMD [19] implementing the STRIDE [20] and DSSP [21] algorithms). In MH^2C each helix is first subjected to a principal component analysis (PCA), yielding three principal components PC1, PC2, and PC3. These are used as a local coordinate system ("reference frame" of the respective helix) for least-square approximation of the C_α atom coordinates by two spline functions: f_2 in the plane PC1-PC2 and f_3 in the plane PC1-PC3, as done in our previous investigation [11]:

$$\overrightarrow{c}(z) = \mathbf{c} = \begin{pmatrix} z \\ f_2(z) \\ f_3(z) \end{pmatrix}. \quad (1)$$

Here, we study splines f_2, f_3 with $K = 0$ interior knots, that is, we consider only one single spline segment comprising polynomials $P[m] = \{p : \mathbb{R} \to \mathbb{R} \mid p(x) = \sum_{i=0}^{m} a_i x^i;\ \ a_i \in \mathbb{R},\ i = 0, \ldots, m\}$ of degrees $m = 2, 3, 4$. We refrained from using interior knots and set $= 0$, yielding a total of three models.

2.4. Global Geometric Quantities. The MH^2C software package was used to extract global shape characteristics of the MHC molecule, which are less affected by short-term fluctuations in time, as compared to single helix parameters. Each helix is represented by a spline, and the interhelical area is represented by a surface defined by "rulings" (i.e., straight lines) spanned between corresponding points (opposite to each other) on these splines [9]. We use M rulings ($1200 \leq M \leq 1500$) parameterized by a common parameter u.

2.4.1. Interhelical Distance and Area of Interhelical Surface. Rulings between splines of the two α-helices $\mathbf{c}_1$ and $\mathbf{c}_2$ (each assigned identical polynomial degrees m) lend themselves for a straightforward triangulation of the ruled surface [22]. From distances between splines

$$d(u_i) = \left\| \mathbf{c}_2(u_i) - \mathbf{c}_1(u_i) \right\| \quad 1 \leq i \leq M \quad (2)$$

and distances between rulings, the total intrahelical area, A, is computed as outlined in [9]. Likewise, median, quartiles, and extreme values (boxplots) of $d(u_i)$ over time are calculated for each i; see Section 3.1. These graphs provide a rough estimate of changes in width of the intrahelical gap (i.e., the binding cleft) both as a function of helical position and of time.

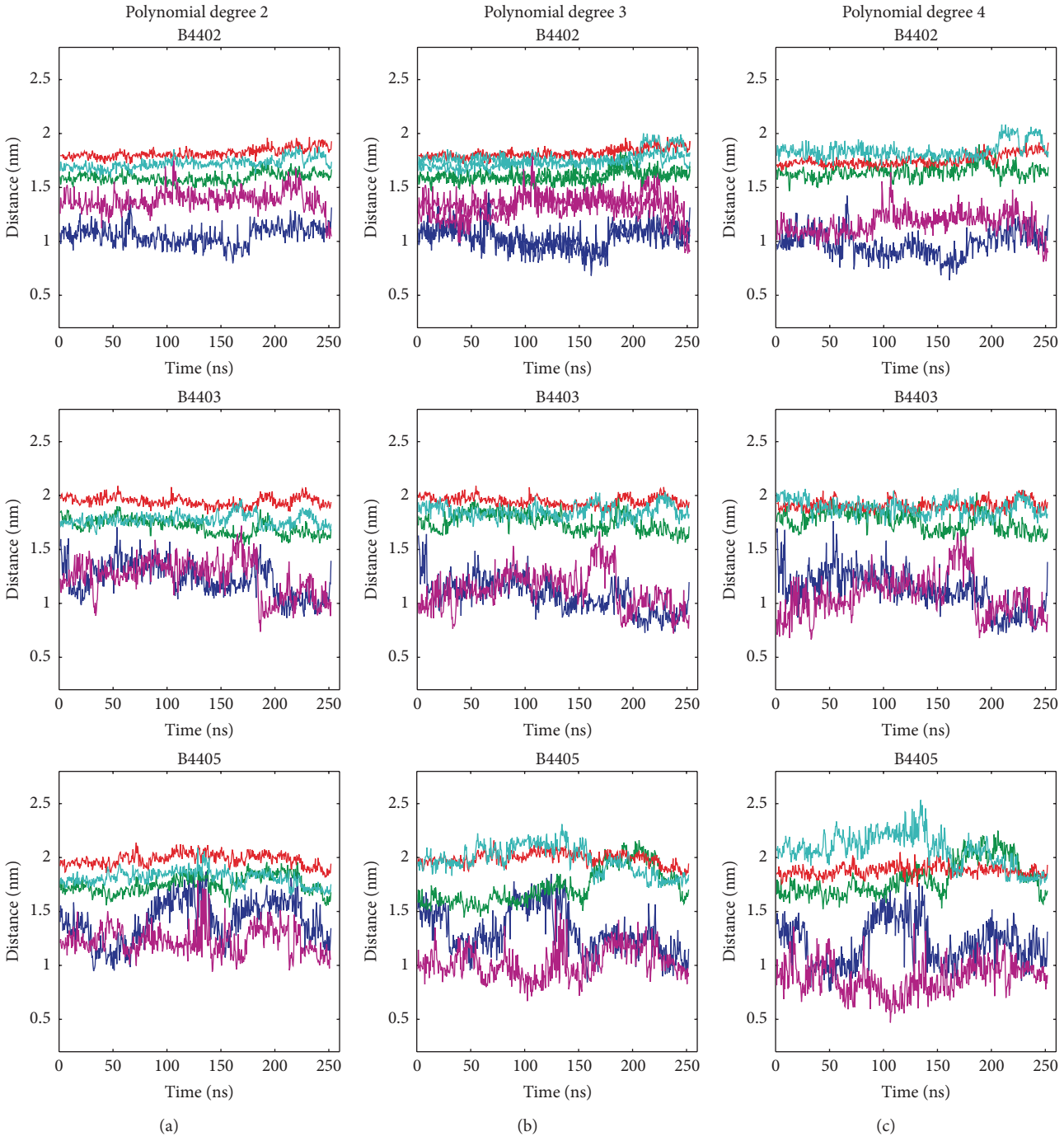


FIGURE 3: *Interhelical distances between MHC α-helices.* Each spline is discretised at about 1500 coordinate points. Interhelical distances between spline positions 1, 369, 737, 1105, and 1471 (blue, green, red, cyan, and magenta, resp.,) were evaluated along a 250 ns MD trajectory for three different molecular systems (B4402, B4403, and B4405) and three different polynomial degrees ($m = 2$, (a); $m = 3$, (b); and $m = 4$, (c)).

3. Results

Three spline models of different polynomial degrees were applied to fit the MHC α-helices of three different molecular systems yielding a total of nine time series per global quantity (see Figures 3 and 5). From the graphs we get an impression of how interhelical distances and the total intrahelical area, A, are affected by different polynomial degrees of the spline functions f_2 and f_3.

3.1. Interhelical Distances. Interhelical distances were measured between five selected points on the splines fitted to G-ALPHA1 helix and G-ALPHA2 helix for polynomial degrees

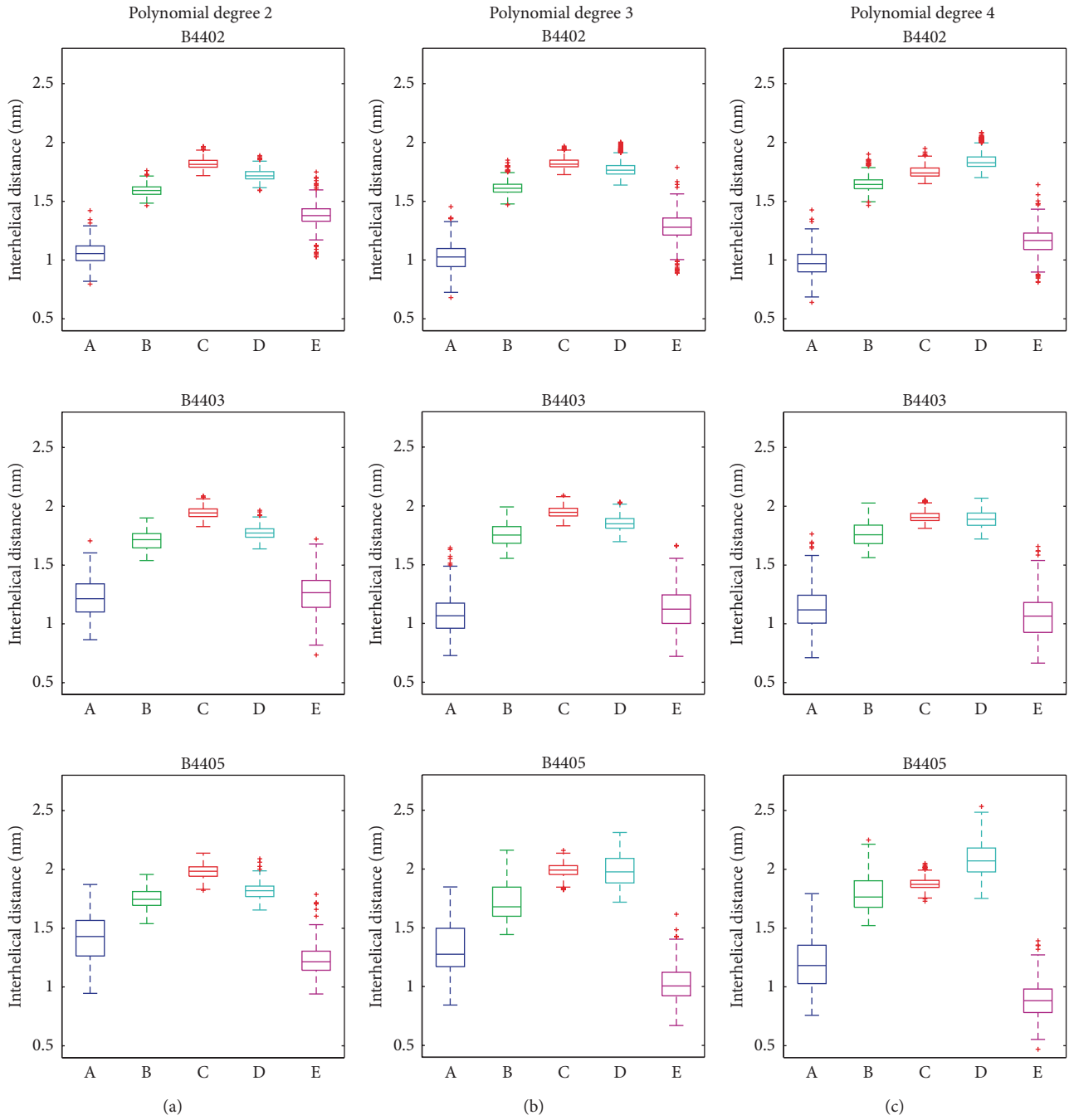


FIGURE 4: *Boxplots of interhelical distances between MHC α-helices.* Each spline is discretised at 1500 coordinate points. Boxplots of interhelical distances between spline positions 1, 369, 737, 1105, and 1471 (blue, green, red, cyan, and magenta, resp., A, B, C, D, and E) are shown along a 250 ns MD trajectory for three different molecular systems (B4402, B4403, and B4405) and three different polynomial degrees ($m = 2$, (a); $m = 3$, (b); and $m = 4$, (c)).

$m = 2, 3, 4$. Each spline was discretised at 1500 discrete coordinate positions from which 1, 369, 737, 1105, and 1471 were selected to describe one aspect of the global shape of the MHC's helical interface (others could include, for example, spline curvature and torsion). Positions 1 and 1471 represent the spline ends or flanking points. Positions 269, 737, and 1105 represent three points of the central part of the splines. On the one hand, in all simulations, the flanking points' distances show the largest fluctuations, but on average are smaller than at the other positions. This reflects the helical bending as seen in Figure 1, panel A. On the other hand, the central parts of the splines show rather little motility for complexes B4402 and B4403 across all models, as seen in Figure 3. For B4405, the central points at positions 369 and 737 show larger fluctuations for models with polynomial degrees 3 and 4 than for the model with degree 2; see Figure 4.

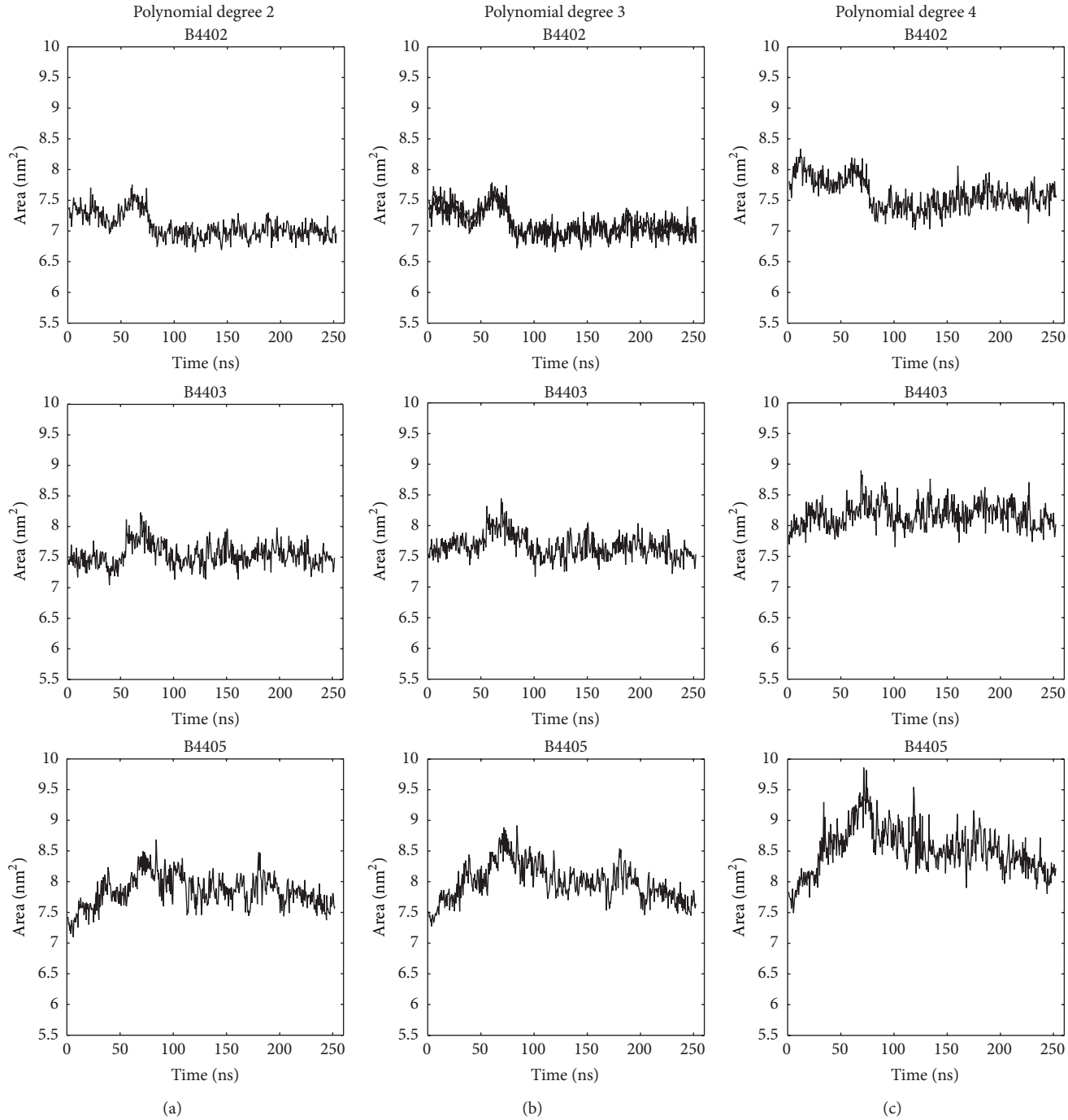


Figure 5: *Area of ruled surface between MHC α-helices*. From distances between splines and distances between rulings, the total intrahelical area, A, is computed along a 250 ns MD trajectory for three different molecular systems (B4402, B4403, and B4405), and three different polynomial degrees ($m = 2$, (a); $m = 3$, (b); and $m = 4$, (c)).

3.2. Area of Ruled Surface between MHC α-Helices. The area of the ruled surface between MHC α-helices, A, was measured between splines fitted to G-ALPHA1 helix and G-ALPHA2 helix for polynomial degrees $m = 2, 3, 4$ for three different MD simulations. The time course of area A is similar for polynomial degrees 2 and 3 (see Figure 5). However, polynomial degree 4 shows an increase of the time averaged area, A: 6.7% for B4402, 7.5% for B4403, and 7.6% for B4405.

4. Discussion

The restriction of T cells to host MHCs is the key mechanism preventing autoimmunity. However, a different shape of an MHC might trigger an immune reaction, even when loaded with self-peptide. Hence, it is of focal interest to spot such changes in MHC geometry, for which we have proposed our geometrical modelling [9]. In the current work, we present a

pilot study on geometrical quantities related to the interhelical area that directly interacts with the T cell receptor that could induce signal transduction.

The new approach in this work is to consider not only single static configurations of MHCs, but to include dynamics. Each specific MHC changes its shape continuously, due to thermal movement. However, these differences in shape do not trigger restricted TCRs. There need to be differences shining through all these thermal movements and becoming relevant in the long MD simulation run.

In our previous work model flexibility was investigated for splines fitted to single helices, which, by their nature, exhibit rather large fluctuations due to motions of small groups of atoms. Here, we investigate model flexibility related to geometric quantities (i.e., interhelical distances and total area between helices) which by their nature resemble more global features of a molecule and should be less affected by stochastic motions of small groups of atoms. Increasing flexibility seems to add short term fluctuations. The question is if these correspond to actual movements of the helices, which are relevant for interpretation, or if they are just artefacts of overfitted models. The impact of increased model flexibility was investigated for rather insensitive quantities such as interhelical distances and total area between MHC α-helices.

Our goal is to detect differences in the dynamics of the TCR/pMHC complexes originating from different MHC molecules (HLA-alleles B*44:02, B*44:03, and B*44:05). In order to achieve that we need to select a reliable model that fits the data accurately and reflects helical motions. As described in our previous work, a reliable model can be selected by the Akaike information criterion [11]. However, automation of model selection holds the risk that different spline models will be selected for different MD simulations. How can we know if variability between different MD simulations is really associated with the different natures of the TCR/pMHC complexes or if the variability results from different spline models? In order to prevent the helix representation from adapting configurations that would not make sense from a physicochemical point of view, one should use polynomials of low degrees. In our case, high polynomial degrees would result in fitting the helix turns that would affect our interpretation of global parameters as well as differential geometric parameters.

For the area, A, of the ruled surface between MHC α-helices, the models with polynomial degrees 2 and 3 yield rather similar mean values. However, the model with degree 4 shows a roughly 7% increase in A, when compared to lower polynomial degrees. Since A is sensitive to model selection, one has to be careful when comparing mean values of A across different simulations. However, the shape of the time series of A is similar across all models, indicating that A is a quantity rather insensitive to small fluctuations during an MD simulation. For A we suggest to use the same model when one wants to compare A between different simulations of similar molecular systems.

For the other quantity described in this work, the interhelical distances, the situation is different; for example, the interhelical distances are comparable between all models for B4403 (Figure 4, second row). However, the same distances show large variations and fluctuations between all models for HLA-B*44:05 (Figure 4, third row). Therefore, we suggest performing a careful analysis before comparing values across different MD simulations. In future studies one could further evaluate the helical dynamics and see if the range of helical structures is well preserved or rather transient in nature over time.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The underlying mathematical concept has been described in [9, 11]. The software implementation was carried out by B. Knapp and B. Hischenhuber. We gratefully acknowledge the support by B. Hischenhuber in implementing the mathematical concept and the corresponding software to novel computed quantities presented in this work. The MD trajectories used in the present work were generated on the IBM-BlueGene computer facility at Bulgarian National Centre for Supercomputing Applications (NCSA). The work was supported in part by BSF and OeAD under Grants nos. DCVP 02/1/2009, DNTS-A 01-2/2013, and WTA-BG 06/2013.

References

[1] S. C. Jameson, K. A. Hogquist, and M. J. Bevan, "Positive selection of thymocytes," *Annual Review of Immunology*, vol. 13, pp. 93–126, 1995.

[2] R. M. Zinkernagel and P. C. Doherty, "Restriction of *in vitro* T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system," *Nature*, vol. 248, no. 5450, pp. 701–702, 1974.

[3] D. Gordon, R. Chen, and S.-H. Chung, "Computational methods of studying the binding of toxins from venomous animals to biological ion channels: theory and applications," *Physiological Reviews*, vol. 93, no. 2, pp. 767–802, 2013.

[4] G. Bao, "Mechanics of biomolecules," *Journal of the Mechanics and Physics of Solids*, vol. 50, no. 11, pp. 2237–2274, 2002.

[5] R. Lavery, A. Lebrun, J.-F. Allemand, D. Bensimon, and V. Croquette, "Structure and mechanics of single biomolecules: experiment and simulation," *Journal of Physics Condensed Matter*, vol. 14, no. 14, pp. R383–R414, 2002.

[6] Y. Cui, "Using molecular simulations to probe pharmaceutical materials," *Journal of Pharmaceutical Sciences*, vol. 100, no. 6, pp. 2000–2019, 2011.

[7] S. A. Adcock and J. A. McCammon, "Molecular dynamics: survey of methods for simulating the activity of proteins," *Chemical Reviews*, vol. 106, no. 5, pp. 1589–1615, 2006.

[8] W. A. Macdonald, Z. Chen, S. Gras et al., "T cell allorecognition via molecular mimicry," *Immunity*, vol. 31, no. 6, pp. 897–908, 2009.

[9] B. Hischenhuber, H. Havlicek, J. Todoric, S. Höllrigl-Binder, W. Schreiner, and B. Knapp, "Differential geometric analysis of alterations in MH α-helices," *Journal of Computational Chemistry*, vol. 34, no. 21, pp. 1862–1879, 2013.

[10] B. Hischenhuber, H. Havlicek, J. Todoric, S. Höllrigl-Binder, W. Schreiner, and B. Knapp, "Corrigendum: differential geometric analysis of alterations in MH α-helices," *Journal of Computational Chemistry*, vol. 34, no. 32, p. 2834, 2013.

[11] B. Hischenhuber, F. Frommlet, W. Schreiner, and B. Knapp, "MH^2c: characterization of major histocompatibility α-helices—an information criterion approach," *Computer Physics Communications*, vol. 183, no. 7, pp. 1481–1490, 2012.

[12] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams et al., "The protein data bank: a computer based archival file for macromolecular structures," *The European Journal of Biochemistry*, vol. 80, no. 2, pp. 319–324, 1977.

[13] J. S. Fraser, M. W. Clarkson, S. C. Degnan, R. Erion, D. Kern, and T. Alber, "Hidden alternative structures of proline isomerase essential for catalysis," *Nature*, vol. 462, no. 7273, pp. 669–673, 2009.

[14] J. Robinson, J. A. Halliwell, H. McWilliam, R. Lopez, P. Parham, and S. G. E. Marsh, "The IMGT/HLA database," *Nucleic Acids Research*, vol. 41, no. 1, pp. D1222–D1227, 2013.

[15] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GRGMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008.

[16] H. J. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, "Interaction models for water in relation to protein hydration," in *Intermolecular Forces*, pp. 331–342, Springer, Amsterdam, The Netherlands, 1981.

[17] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.

[18] U. Omasits, B. Knapp, M. Neumann et al., "Analysis of key parameters for molecular dynamics of pMHC molecules," *Molecular Simulation*, vol. 34, no. 8, pp. 781–793, 2008.

[19] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[20] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins: Structure, Function and Genetics*, vol. 23, no. 4, pp. 566–579, 1995.

[21] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[22] M. Peternell, H. Pottmann, and B. Ravani, "On the computational geometry of ruled surfaces," *Computer-Aided Design*, vol. 31, no. 1, pp. 17–32, 1999.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 871810, 11 pages
http://dx.doi.org/10.1155/2014/871810

Research Article

A Mathematical Model of Skeletal Muscle Disease and Immune Response in the *mdx* Mouse

Abdul Salam Jarrah,[1] Filippo Castiglione,[2] Nicholas P. Evans,[3] Robert W. Grange,[4] and Reinhard Laubenbacher[5,6]

[1] *Department of Mathematics and Statistics, American University of Sharjah, Sharjah 26666, UAE*
[2] *Institute for Applied Mathematics, National Research Council of Italy, 00185 Rome, Italy*
[3] *Department of Population Health Sciences, Virginia Tech, Blacksburg, VA 24060, USA*
[4] *Department of Human Nutrition, Foods and Exercise, Virginia Tech, Blacksburg, VA 24060, USA*
[5] *Center of Quantitative Medicine, University of Connecticut Health Center, Farmington, CT 06030, USA*
[6] *Jackson Laboratory for Genomic Medicine, Farmington, CT 06030, USA*

Correspondence should be addressed to Abdul Salam Jarrah; ajarrah@aus.edu

Received 30 March 2014; Accepted 19 May 2014; Published 11 June 2014

Academic Editor: Francesco Pappalardo

Duchenne muscular dystrophy (DMD) is a genetic disease that results in the death of affected boys by early adulthood. The genetic defect responsible for DMD has been known for over 25 years, yet at present there is neither cure nor effective treatment for DMD. During early disease onset, the *mdx mouse* has been validated as an animal model for DMD and use of this model has led to valuable but incomplete insights into the disease process. For example, immune cells are thought to be responsible for a significant portion of muscle cell death in the *mdx* mouse; however, the role and time course of the immune response in the dystrophic process have not been well described. In this paper we constructed a simple mathematical model to investigate the role of the immune response in muscle degeneration and subsequent regeneration in the *mdx* mouse model of Duchenne muscular dystrophy. Our model suggests that the immune response contributes substantially to the muscle degeneration and regeneration processes. Furthermore, the analysis of the model predicts that the immune system response oscillates throughout the life of the mice, and the damaged fibers are never completely cleared.

1. Background

Duchenne muscular dystrophy (DMD) is a lethal, X-chromosome muscle wasting disease affecting approximately one in 3,500 boys [1, 2]. Patients appear clinically normal at birth with the exception of elevated serum creatine kinase levels. The onset of DMD begins in early childhood with the first observed symptoms between two and five years of age. Typically by the age of 12, DMD patients require the use of a wheelchair due to the loss of lower limb muscle strength. Progressive weakness of the arms and legs, along with kyphoscoliosis, continues through late disease progression. Many patients die in their late teens or early twenties due to respiratory or cardiac complications [1, 3]. Currently, there are no effective means of therapy or treatment for DMD.

In 1984, Bulfield et al. identified a spontaneous mutation in C57BL/10ScSn inbred mice that exhibited a disease state similar to human DMD [4]. The X chromosome-linked mutation resulted in mice (*mdx* mice) with high serum levels of muscle enzymes and with histological lesions comparable to those seen in human muscular dystrophy. This mutation in the murine dystrophin gene caused an absence of dystrophin in skeletal muscle and this key defect validated the *mdx* mouse as a suitable model of the early onset of DMD human disease [5, 6].

The histology and time course of the disease in *mdx* mouse model are very different from those in DMD patients: relatively normal life span and overall fitness compared to progressive physical impairment leading to death in DMD patients [7]. Nonetheless, the *mdx* mouse model is regarded

as the best animal model, especially of the early onset of DMD.

Dystrophin deficiency does not always produce muscle degeneration at all life stages, in all muscle phenotypes, or in all animal models [8]. In dystrophin-deficient skeletal muscle, for example, mechanical injury and proteolysis may be important factors but do not fully explain DMD pathogenesis. Mechanisms such as the immune/inflammatory response to injury appear to contribute substantially to muscle pathophysiology. Observations of activated immune cell infiltrates in dystrophic muscle suggest that the immune/inflammatory response may play a role in exacerbating the disease [8–12].

Immune/inflammatory-mediated mechanisms, which result in muscle cell death and/or mechanisms leading to fibrosis, may be important initiators of lesions in dystrophin-deficient muscle. Large populations of lymphocytes, macrophages, and neutrophils are present in DMD muscle tissue [10]. T-cells and macrophages are classically thought to be responsible for triggering and orchestrating the immune response, inducing target cell death, recognizing immune stimuli, and removing cellular debris. Immunosuppressive therapy, such as treatment with glucocorticoids, improves muscle strength and prolongs ambulation in DMD patients but does not prevent disease progression [3, 12]. The article [13] is a comprehensive literature review of the immune-mediated molecular and signaling mechanisms that regulate the time course of the disease and the *mdx* mouse model.

One limitation found in the DMD literature is that there are few time course datasets that are consistent for muscle type, ages, or collected with the same methods. To our knowledge, there is not a time course dataset that accounts for the aspects of the immune response, muscle degeneration, and muscle regeneration as modeled herein.

In this paper we presented a simple mathematical model to investigate the role of the immune response in muscle degeneration in *mdx* mice. The mathematical model represents a novel approach to study DMD pathogenesis and to identify potential therapeutic targets. Using the available data, we constructed a mechanistic differential equations model as the first step toward building a comprehensive model of the immune response in DMD aiming to provide insight into the nature of the immune/inflammatory mechanisms contributing to DMD pathogenesis in the early disease stages. Understanding these underlying mechanisms will provide a key tool to develop effective therapeutic approaches.

The model incorporates the generally held hypothesis that the immune response contributes to muscle tissue damage via CD8+ T-cells that are recruited by macrophages through CD4+ T-cells. This simple model fits the available experimental time course data found in the literature. Moreover, the model suggests that CD8+ T-cells likely contribute to muscle damage and predicts two distinct modes for the long-term dynamics of the immune response.

2. Experimental Techniques and Data

Immune and histopathological time course data used for the model were obtained from available literature. We combined time course data for several different dystrophic muscles for the age range 14–84 days, including those for the concentrations of CD8+ and CD4+ T-cells in quadriceps [9], macrophages in soleus [11], procion orange dye uptake (as an indicator of fiber damage) in *soleus (SOL)* and histopathology in soleus and extensor *digitorum longus* (EDL) muscles [15], and TA muscles [14]. It should be noted here that the *extensor digitorum longus* (EDL) and *tibialis anterior* (TA) muscles are primarily fast twitch muscles whereas the *soleus* (SOL) is a slow twitch muscle.

Throughout this paper, we define *normal* muscle fibers as those that do not exhibit damage; that is, they are neither degenerating nor necrotic, nor do they demonstrate uptake of procion orange dye. *Damaged* fibers are those that are degenerating or necrotic. *Regenerating* fibers are those that were damaged and undergoing repair (identified by the presence of centralized nuclei).

3. The Mathematical Model

The model presented here describes in mathematical terms the action of the immune system on muscle tissues subsequent to damage triggered by a not well-specified mechanism. Since we are interested in following the role of the immune system during disease progression, we use a simplistic description of the development of damaged tissue while we model the cell interaction during the buildup of the adaptive (i.e., specific) immune response in more detail. In particular, we include macrophages that trigger T-helper activation enabling cytotoxicity for muscle cells. Whereas macrophages are employed in the removal of damaged cells, favoring tissue regeneration, they also stimulate specific CD8+ lymphocytes that in turn create additional damage. The model is based on the following set of variables: the concentration of immune cells (CD4+, CD8+ T-cells, and macrophages) and the fractions of morphologically normal muscle fibers, damaged muscle fibers, and regenerating muscle fibers in the muscle tissue.

The model variables are immune cells (cell numbers in a cubic millimeter of muscle tissue), macrophages (M) CD4+ T-helper lymphocytes (H) and CD8+ cytotoxic T lymphocytes (C), and *muscle fibers* (percentage of the whole muscle tissue): normal (N), damaged (D), and regenerating fibers (R).

The model equations are as follows:

$$\frac{dH}{dt} = b_H + k_1 DM - d_H H, \tag{1}$$

$$\frac{dC}{dt} = b_C + k_2 DH - d_C C, \tag{2}$$

$$\frac{dM}{dt} = b_M + k_3 MD - d_M M, \tag{3}$$

$$\frac{dN}{dt} = k_4 R - k_5 CN - \alpha N, \tag{4}$$

$$\frac{dD}{dt} = k_5 CN + \alpha N - k_6 DM - d_D D, \tag{5}$$

$$\frac{dR}{dt} = k_6 DM + d_D D - k_4 R. \tag{6}$$

Here $\alpha = \alpha(t)$ represents mechanical damage as a lognormal function of time. This damage triggers the cascade of events starting with macrophage infiltration into the tissue, recruitment of CD4+ cells, and activation of CD8+ cells. The choice of a lognormal function to model muscle cell failure comes from considering the accumulation of damage in the muscle tissue as a multiplicative degradation process. In brief, the degradation process leading to a lognormal model is obtained when the amount of degradation or damage that ends in complete failure follows the relationship $y_t = (1 - \varepsilon_t)y_{t-1}$, where ε_t are small independent random shocks (in our case occurring to each single muscle cell) during progression to failure. This relationship means that the increase in the amount of damage from one time point to the next is a small multiple of the total amount of damage accumulated (i.e., fewer undamaged muscle cells have to share the same muscle workload, resulting in a larger effort or stress that yields increased damage at the next time point). When a large number of cells are damaged, by the central limit theorem, the probability of failure can be approximated by a lognormal distribution, or, to be more precise, since failure occurs when the amount of degradation reaches a critical point, the time to failure will be modeled by a lognormal [16],

$$\alpha(t) = \frac{h}{t\sigma\sqrt{2\pi}} e^{-(\ln(t)-m)^2/(2\sigma^2)}, \quad (7)$$

where m, h, and σ are parameters to be determined on the basis of the available experimental data.

For this model, we assume that the physiological damage arising from regular muscular activity is responsible for the initial mechanical damage. Due to genetics and other factors that are determinant in developing the disease, this initial damage is amplified in the *mdx* mice leading to a strong immune response, while in the wild-type mice (disease-free) the initial damage will not be amplified resulting in a balanced immune response. We model this assumption by simply assuming that, in the wild-type model, the amplitude h, which is proportionally related to the mechanical damage, is negligible with respect to its counterpart in *mdx* mice.

Equations (1)–(3) represent the rate of change of the immune cell counts of CD4+ (H) and CD8+ T cells (C) and macrophages (M), respectively. In particular, (2) shows that lymphocytes CD8+ T-cells are replenished at a constant rate b_C and die at the rate $d_C C$. The term $k_2 DH$ stands for the activation of CD8+ cells from damaged cells, but only in presence of helper T-cells which activate them. Similarly the term $k_1 DM$ in (1) describes the activation of CD4+ T-cells by macrophages in the presence of damage. Alternatively macrophages in (3) may also be activated by the damage alone (i.e., do not require a second signal).

Equations (4)–(6) represent the rate of change of the percentage of morphologically normal (N) damaged (D) and regenerating (R) muscle fibers. Equation (5) reflects the assumption that the damage accumulates according to the multiplicative degradation process described above with a time-to-failure $\alpha = \alpha(t)$ modeled as a lognormal (term $-\alpha N$ in (4), and the respective term αN in (5)).

Damaged cells are either removed by the macrophages at rate $k_6 DM$ or by other mechanisms (unspecified) at the rate $d_D D$. Equation (4) indicates that normal cells are replenished from regenerating fibers at the rate $k_4 R$ and are damaged at the rate $k_5 CN$ by the presence of specific CD8+ T-cells.

Note that (6) can be rewritten as $dR/dt = -(dN/dt + dD/dt)$, which is the assumption that a muscle fiber can be either normal, damaged, or regenerating while the total number of fibers in a muscle remains constant.

The model assumes that if there are no damaged muscle fibers in the tissue, the numbers of different immune cells (i.e., CD8+, CD4+ T-cells, and macrophages) do not change. Furthermore, the assumption that the muscle tissue initially has no damaged or regenerating fibers implies the following initial baselines: $N(0) = 100$ and $D(0) = R(0) = 0$.

3.1. Parameter Estimation. In addition to the cell counts at time zero for the three types of immune cells, there are 13 parameters in the model. As nothing is known about the values of these parameters in the literature, especially in a muscle tissue, we estimated their values by fitting, using the freely available software *COPASI* [17]. COPASI is a stand-alone software for the simulation and analysis of network models and their dynamics. It has many functions, including parameter estimation and sensitivity analysis, both of which we used throughout this study. The optimal parameter values presented in Table 1 were obtained by using particle swarm and the genetic algorithms to search for a global solution. Hooke-Jeeves and Lavenberg-Marquardt algorithms were used to find the best local solution. All of these and many other known optimization algorithms are already implemented in COPASI. We used the immune response experimental data from [9, 11] as well as the muscles data from [14, 15] to estimate the parameters of the model.

We assume that the cell count of any type of the immune cells at time zero in *mdx* mice is equal to the cell count of that type of immune cell in a wild-type mouse. Based on the available data in the literature about wild-type C57 mice [15], we assumed that there are initially no CD4+ T-cells, and we set the initial number of macrophages in the tissue to 400 cells and the number of CD8+ T-cells to 4. It should be noted here that the model is not sensitive, however, to any of the initial counts of the immune cells, as we will explain in the next section.

3.2. Model Assessment. Even though the model is simple and includes only the key mechanisms of the immune system, it captures the main features of the known dynamics of *mdx* dystrophic muscle pathophysiology, and it manages to reproduce the seemingly unrelated experimental datasets that are reported in the literature as shown in Figure 1. The trajectories of the immune cells in the left panel of Figure 1 fit well the known immune response time course data about macrophages in soleus from [11] and the CD4+ and CD8+ T-cells from [9], while at the same time, the figures in the right panel of Figure 1 show that the model trajectories of the different muscle tissue types are in good agreement with the SOL time course from [15] and the TA time course from [14].

The EDL time course data [15] was not used in the model calibration. However, the model trajectories follow the same

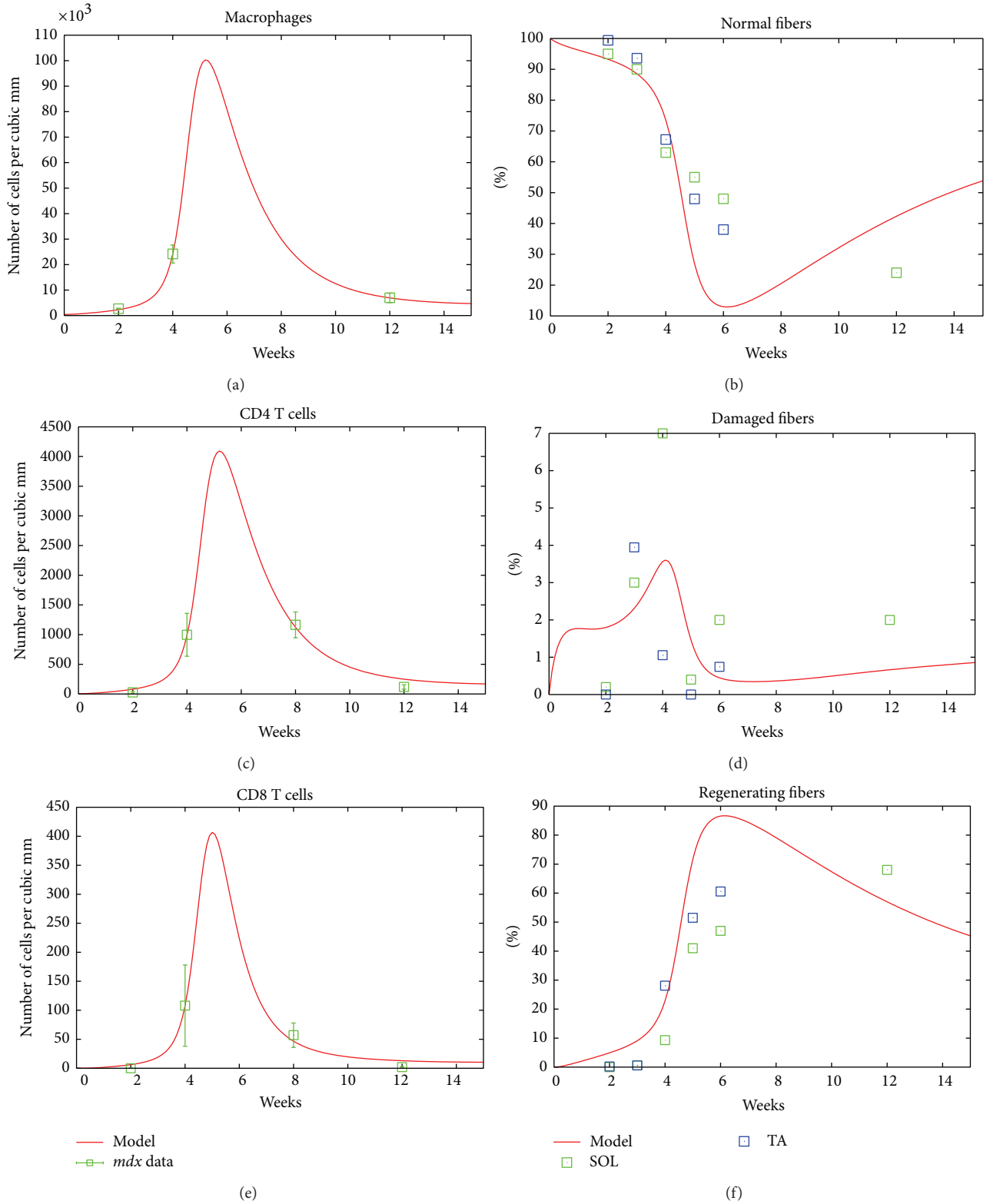


FIGURE 1: *mdx* mice. The left panel shows the fit of the simulated model to the immune response data from [11]. The right panel shows the fit to muscle tissue damage data to soleus muscle taken from [9] and the TA [14].

Table 1: The model parameters and their estimated numerical values (in 1 mm^3 of muscle tissue). Here "c" is for cell and "w" is for week.

Parameter	Description	Value	Unit
b_H	Turnover rate of CD4+ T cells	$d_H H_0$	$c\,w^{-1}$
k_1	Damage-driven proliferation rate of CD4+ T cells	0.0324139	$(cent)^{-1}\,w^{-1}$
d_H	Death rate of CD4+ T cells	0.83355	w^{-1}
b_C	Turnover rate of CD8+ T cells	$d_C C_0$	$c\,w^{-1}$
k_2	Damage-driven proliferation rate of CD8+ T cells	0.115375	$(cent)^{-1}\,w^{-1}$
d_C	Death rate of CD8+ T cells	1.61511	w^{-1}
b_M	Turnover rate of macrophages	$d_M M_0$	$c\,w^{-1}$
k_3	Infiltration rate of macrophages	0.766576	$(cent)^{-1}\,w^{-1}$
d_M	Death rate of macrophages	0.781155	w^{-1}
k_4	Generation rate of healthy fibers	0.123848	w^{-1}
k_5	Cytotoxicity degradation rate	4.09948×10^{-3}	$c^{-1}\,w^{-1}$
k_6	Cleaning rate by macrophages	3.23097×10^{-4}	$c^{-1}\,w^{-1}$
d_D	Physiological cleaning rate	1.34671	w^{-1}
σ	Standard deviation of the initial damage	2.92815	
m	The time of the peak of the initial damage	4.22686	
h	Proportional to the magnitude of the damage	0.511657	
H_0	The initial number of CD4 T cells	0	c
C_0	The initial number of CD8 T cells	4	c
M_0	The initial number of macrophages	400	c

patterns as the EDL data, albeit the EDL dynamics are slower as Figure 3 shows.

As mentioned above, the initial counts of the immune cells in the model are set to be equal to those of the wild-type mice. This assumption, however, might not actually be the case for the *mdx* mouse, as the immune response might have started during gestation (prior to birth). To see the effects of onset of the immune response during gestation on the model curves and to explore the sensitivity of the model to these parameters, we varied the values of the three counts and calculated the new trajectories of the 5 variables of the model. Figure 4 shows the trajectories of macrophages and damaged fibers of 125 different sets of initial values of the three immune cell types. These random values were picked from ranges determined from the available data from wild-type mice. It is clear that these trajectories are qualitatively the same as the corresponding trajectories in Figure 1. This outcome shows that the model is insensitive to the initial counts of immune cells.

3.3. Model Prediction. As shown in Figure 1, our model trajectories are in good agreement with the available *mdx* mice datasets. Furthermore, the wild-type datasets are in good agreement with our model as shown in Figure 2, where the amplitude h of the mechanical damage in the disease-free mice (wild-type) is one tenth of its counterpart in the *mdx* mice.

To test the model sensitivity to its parameters, we varied the parameters' values from within their ranges as reported in Table 2. For each of the listed parameters, two values were chosen from within the range of that parameter. This approach resulted in more than 1000 parameter sets. Figure 5 shows the dynamics of macrophages and normal fibers resulting from a sample of these sets. It is clear in Figure 5 that, for the first 7 weeks, the trajectories are qualitatively the same for all parameter sets. However, for the weeks after, and for most of the parameter sets, we see oscillation both in the immune response as well as the different types of muscle fibers.

Table 2: Some of the model parameters (column 1), their estimated values (column 2), and, for each parameter, a lower (column 3) and an upper (column 4) bound. To study the sensitivity of the model to a given parameter, we varied the value of that parameter within its range above, and qualitatively examined the resulting dynamics.

Parameter	Value	Min	Max
k_1	0.0324139	0.0160773	0.0643092
d_H	0.83355	0.416775	1.6671
k_2	0.115375	0.0577452	0.230981
d_C	1.61511	0.807555	3.23022
k_3	0.766576	0.381371	1.52549
d_M	0.781155	0.390577	1.56231
k_4	0.123848	0.061924	0.247696
k_5	4.09948×10^{-3}	2.04974×10^{-3}	8.19896×10^{-3}
k_6	3.23097×10^{-4}	1.61548×10^{-4}	6.46194×10^{-4}
d_D	1.34671	0.673355	2.69342

The literature suggests that "cycles" of muscle degeneration and regeneration contribute to progressive muscle wasting [18–20] which is likely antagonized further by the immune response [8–12]. This relationship suggests a predator-prey-like interaction between the immune system and the tissue, which has been reported in many other diseases [21]. After the dramatic increase of the immune cells and the initial peak of damage in the first 4–8 weeks, the immune

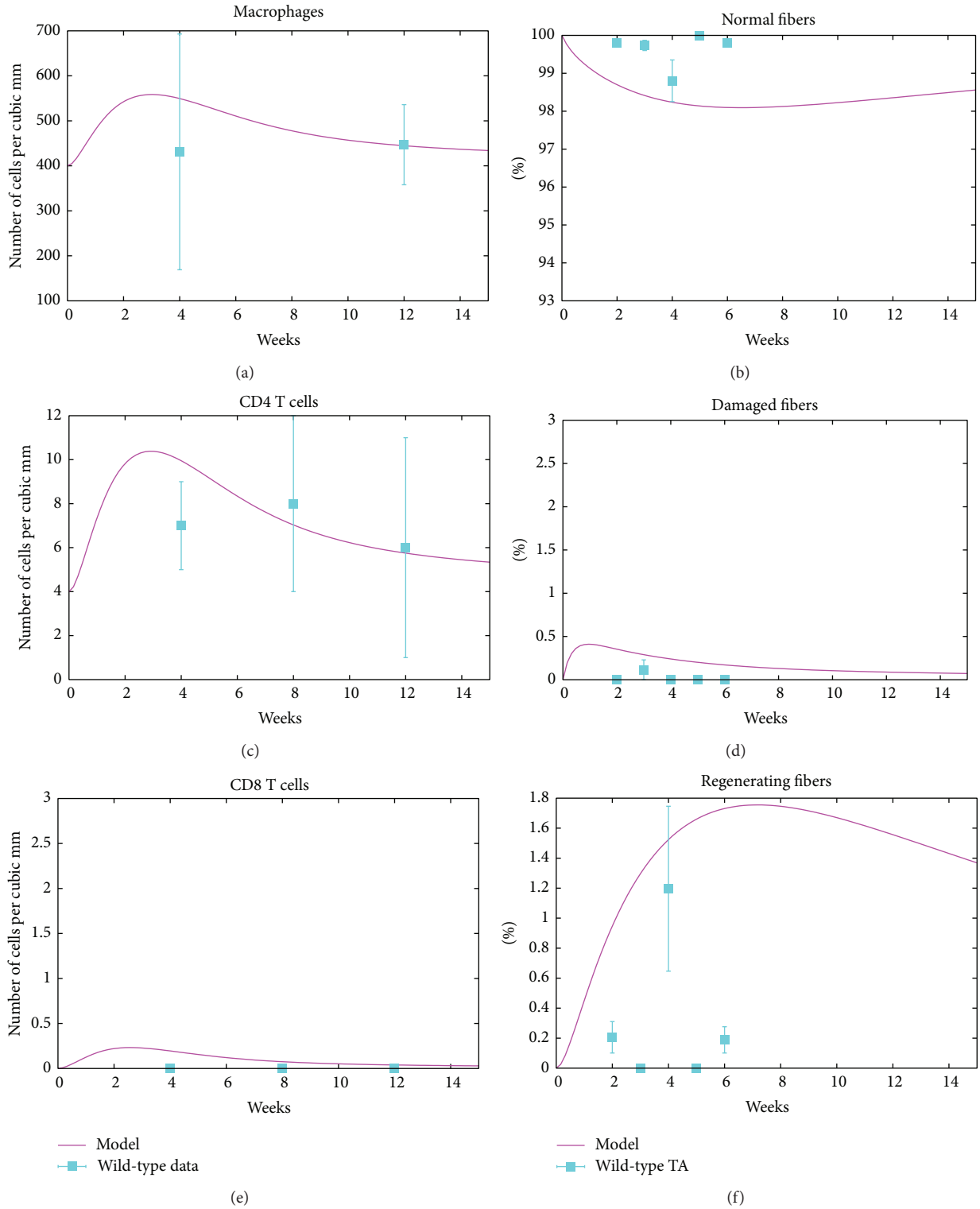


FIGURE 2: Wild-type mice. The left panel shows the fit of the simulated model to the immune response data from [11]. The right panel shows the fit to muscle tissue data from TA [14].

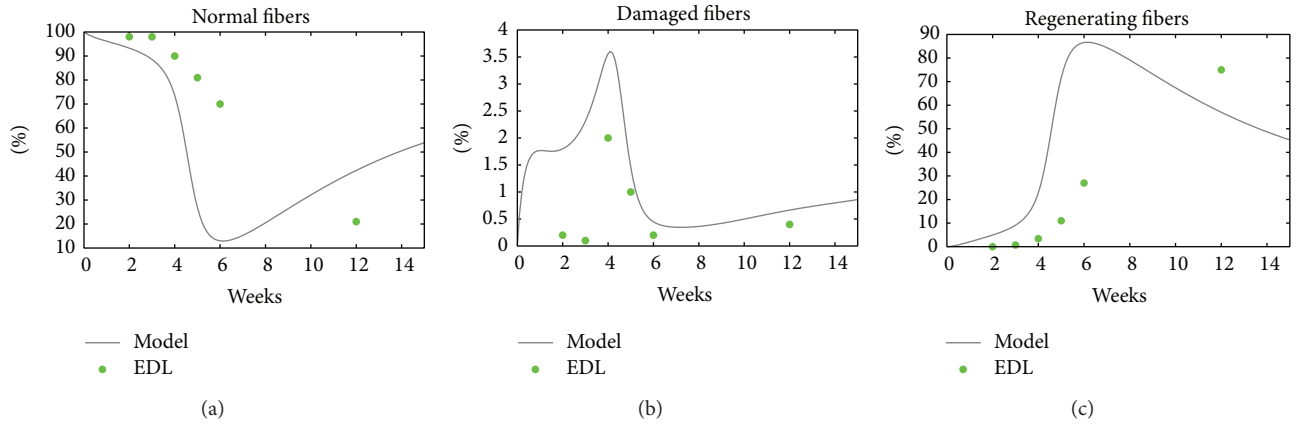


Figure 3: EDL data and the model curves. The model trajectories and the data follow the same pattern, even though the disease time course seems to be slower in EDL than in our model, and this could result because EDL muscle has different physical characteristics from the SOL at TA muscles that were used to fit the model.

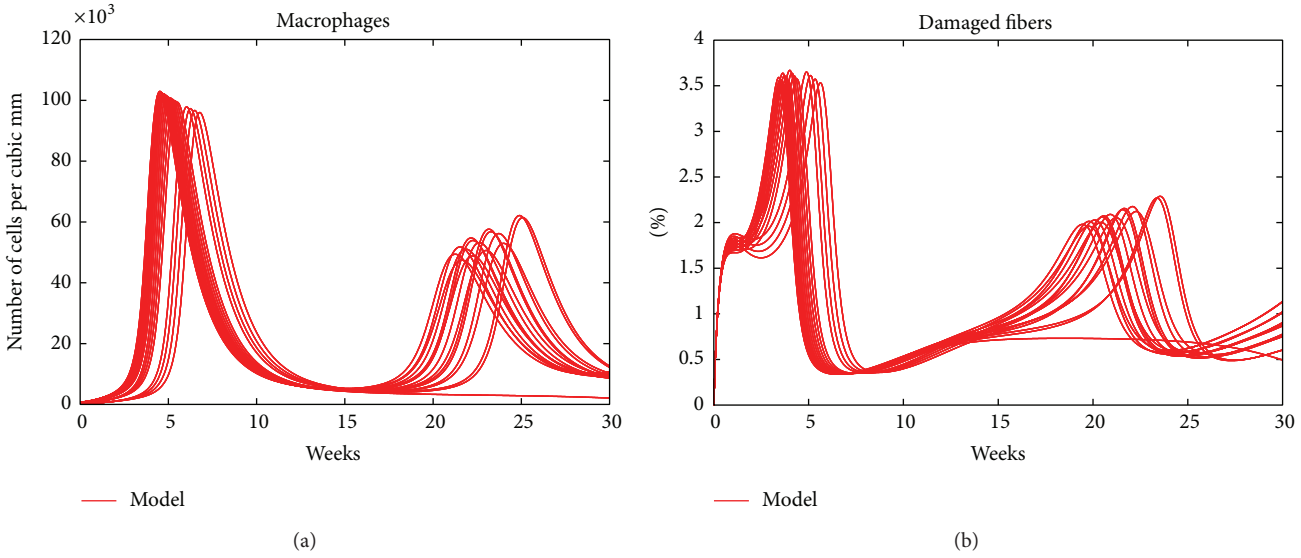


Figure 4: The trajectories of macrophages and damaged fibers starting from 125 different initializations of the immune cell numbers.

cells decrease; muscle degeneration and regeneration cycles are suppressed but continue at low levels. This scenario or mode is captured in Figure 5, by many parameter sets, where oscillation with low amplitude is persistent in the immune response as well as the normal fibers.

For almost all other parameter sets, another mode emerged: the system escapes the oscillatory behavior above and slowly approaches a disease-free state where the percentages of damaged and regenerating fibers will both be very small, and almost all tissue fibers are normal.

This could be the result of a significant decrease of the immune response. In this case, even though regeneration is slow, pathophysiology decreases and degeneration diminishes, which may further suppress the immune response.

The time course for the dystrophic process in *mdx* mice for the first year of life includes muscle degeneration/regeneration cycles between the age of 3 and 10 weeks, followed by a decrease in these cycles and relative muscle stability to the age of ~1 year [19]. The outcomes of the model represented by both modes appear to account for the attenuation of the dystrophic process in mice after the age of ~10 weeks.

Furthermore, the model predicts that the dynamics could switch between the two modes in a response to a change in some of the key parameters, such as the rate at which regenerating fibers become normal.

3.4. Immune Cell Depletions. Elevated concentrations of macrophages (>80,000 cells/mm^3, normal ~1000 cells/mm^3) have been observed in 4–8 weeks *mdx* mouse muscle but rapidly decrease by 12 weeks [9, 11, 15]. Macrophages have a variety of immunoregulatory and inflammatory functions. They are rich sources of cytokines and nitric oxide, a potent-free radical that can lyse muscle cells. Macrophages are

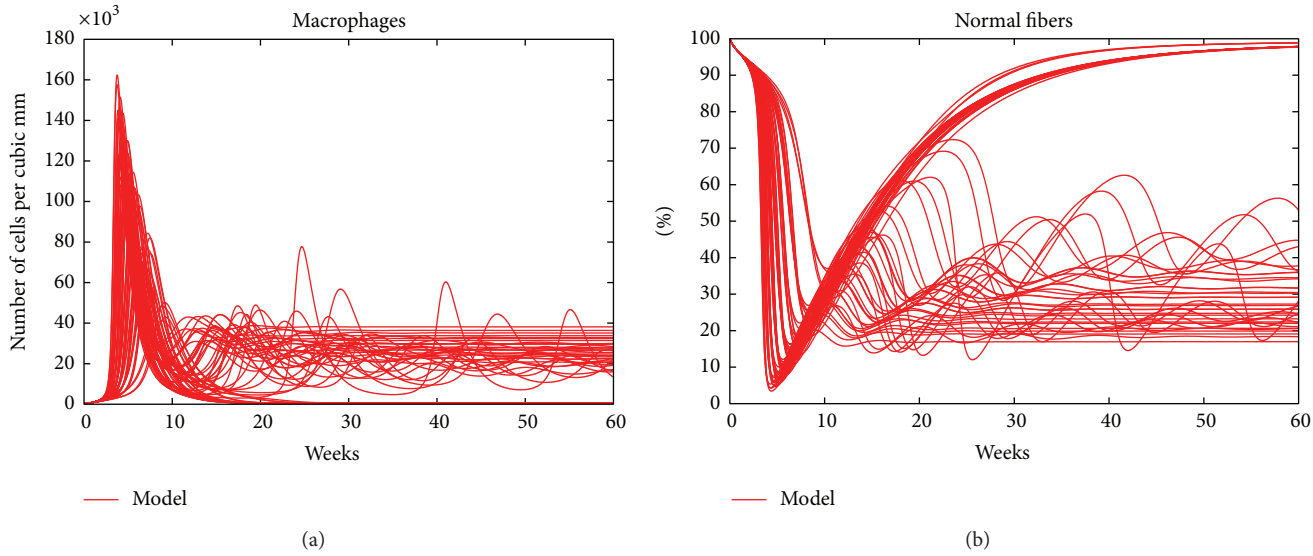


Figure 5: The dynamics of macrophages and normal fibers for a sample of parameter sets from within the specified ranges in Table 2.

antigen-presenting cells, which may regulate the immune response in dystrophic muscle and could possibly activate T-cells. Antibody-mediated depletion of macrophages, beginning at 6 days old and continuing to 4 weeks in *mdx* mice, resulted in a >75% reduction of injured muscle fibers, suggesting an important role of these cells in the development of lesions [11].

Elevated concentrations of activated CD8+ and CD4+ T-cells are observed in affected muscles of *mdx* mice at the age of 4–8 weeks, but rapidly decrease by 14 weeks (>1200 cells/mm^3 in *mdx* muscle, normal ~100 cells/mm^3) [9, 15]. Antibody-mediated depletion of CD8+ or CD4+ T-cells in *mdx* mice, beginning at 6 days old and continuing to 4 weeks of age, resulted in a 75% and 61% reduction in muscle pathology, respectively, for mice at the peak of disease progression [22].

According to the model construction, CD8+ T-cells contribute to damage and are produced mainly in response to increased numbers of macrophages; hence, depleting macrophages should reduce the percentage of CD8+ T-cells in the tissue which will do two things: (i) there will be no damage due to CD8+ T-cells, but it also means that there will be no macrophages to clear the damaged cells; (ii) in the model, the macrophage contribution to clearing out the damage is high and so depleting the macrophages will initially increase the damage but then the damage decreases as shown in Figure 6.

Figure 6(a) shows that macrophage depletion in the model resulted in improved muscle tissue with 86% of the muscle fibers characterized as normal. It should be noted here that the percentage of damaged fibers in Figure 6(b) in the tissue is not much lower than the one in Figure 1(d). This outcome may not be surprising since depleting macrophages yields fewer CD8+ T-cells and hence fewer degenerating fibers. At the same time, however, there will be fewer CD4+ T-cells and hence less stimulus to promote the regenerating process.

4. Discussion and Conclusions

Immune response is known to play a key role in exacerbating the disease in DMD patients and *mdx* mice. Indeed, a primary treatment for DMD patient is glucocorticoids (GC) whose main effect is the inhibition of numerous inflammatory genes. Even though *mdx* mice show a mild phenotype and do not accurately reflect the severe nature of the human disease, they are considered a reasonable model. In *mdx* mice, several studies demonstrated that depletion of immune cells results in improvement of dystrophic muscle pathophysiology; however, the role and time course of the immune response in the dystrophic process have not been well described. Even though there are time course datasets from different muscles that account for different features of the disease, to our knowledge, there is not a single time course that accounts for both the immune response and muscle pathophysiology in the same muscle type.

We have presented a mathematical model of the immune response to the skeletal muscle wasting and inflammation in *mdx* mice. Our model is simple in the sense that it contains only the basic mechanisms of the immune response and its known interactions with dystrophic muscle. Nonetheless, our model reproduces available data from different experimental studies. Furthermore, the depletions of the immune cells in the model result in an increased percentage of normal fibers as observed in different experiments on *mdx* mice.

The dynamics of the tuned system shows one main behavior in which the immune system contributes to damage in recurring phases of muscle destruction. However, by changing some crucial parameters' values (in particular parameters k_1, k_2, and k_5) the frequencies of these relapses can be modified eventually obtaining a mode in which the disease is controlled, where damaged fibers regenerate but slowly. The above three parameters are the damaged-driven proliferation rates of CD4+, CD8+ T-cells, and

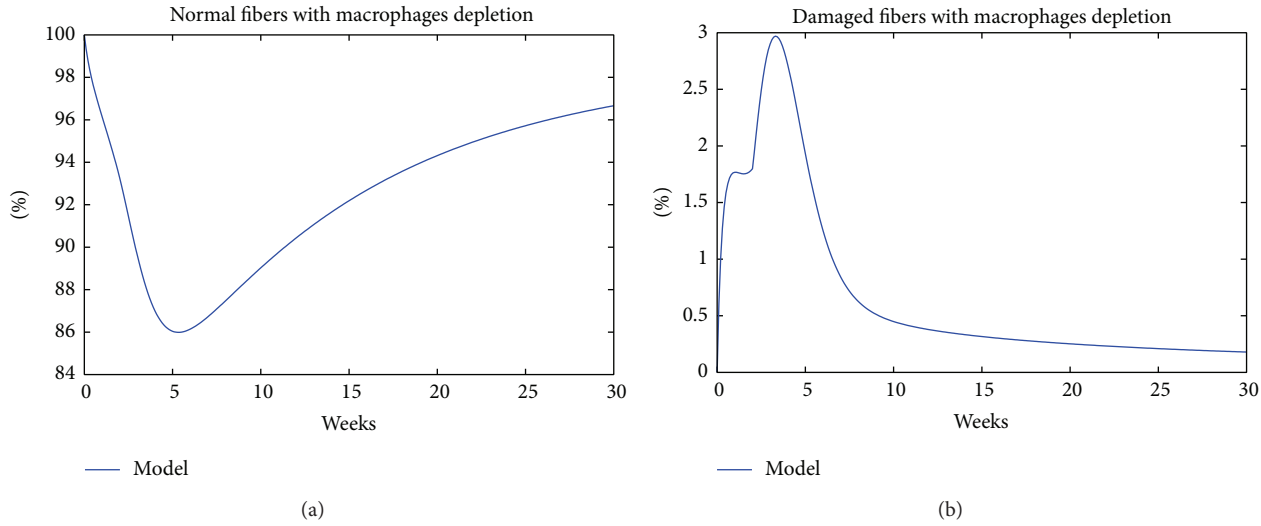


FIGURE 6: Effect of macrophage depletion on normal and damaged fibers, respectively, in (a) and (b).

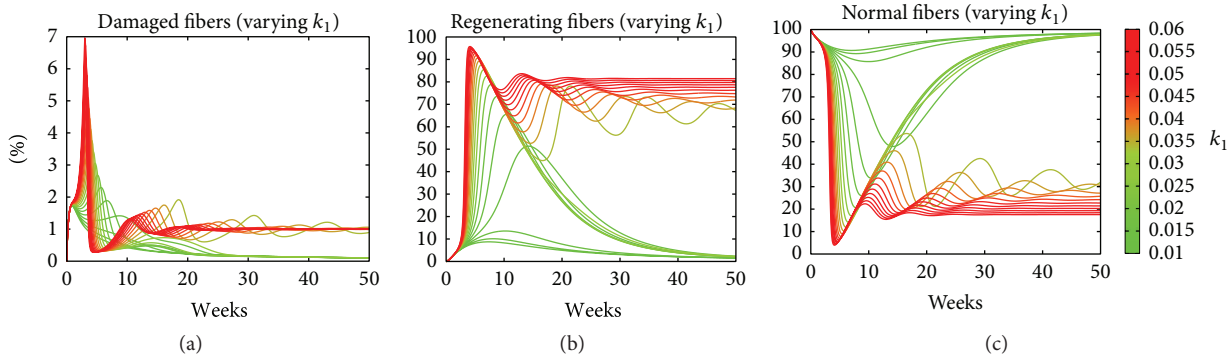


FIGURE 7: The effect of varying the parameter k_1, while fixing all other parameters, on the trajectory of fibers. As the value of k_1 increases, more tissue goes from damaged (panel (a)) to regenerating (b) and back to normal (c). Interestingly the oscillatory behaviour is found somewhere in between of the two regimes corresponding to k_1 at the lower and at the higher edge of the interval values.

the cytotoxicity of CD4+ T-cells, which are key drivers of the damage in the model. Figure 7 shows the dependency on k_1 of the trajectories of damaged, regenerating, and normal fibers. As the value of k_1 increases, both the percentage of damaged fibers and the intensity of the immune response increase, leading to an accumulation of regeneration fibers during the period of 5–7 weeks, and this results in an oscillatory behavior where the damage is persistent and around 75 percent of the tissue is regenerating. Similar curves appear when we performed the same analysis using k_2 or k_5.

The heat maps in Figure 8 show the asymptotic (at week 50) percentages of normal and regenerating when the parameters k_1 and k_2 are changed together. The two main regimes are found for extreme values of the two parameters going from regenerating to normal by increasing both k_1 and k_2. It clearly shows that, by increasing both parameters together, the percentage of normal (regenerating, resp.) would increase (decrease, resp.) at a faster rate than increasing one parameter only. Furthermore, it is easy to see that a small decrease in any of the two parameters would lead to the oscillatory behavior where the majority of the fibers are regenerating.

Another crucial parameter is the rate at which a regenerating fiber returns, in our model, to morphologically normal muscle fiber, k_4, which is low indicating a slow regeneration process. Once a fiber is damaged, it could take weeks before it returns to a morphologically normal muscle fiber. This could explain the high percentage of regenerating fibers observed in dystrophic muscles. The model is sensitive to this parameter. Figure 9 shows how the percentages of damage, regenerating, and normal tissue change when varying k_4. As we increase the value of k_4, the normal tissue asymptotically increases (see Figure 9(c)). However, beyond an optimal value, $k_4 \sim 0.1$ for which the normal cells are about 100% of the total, the percentage of normal tissues falls to and oscillates around 30%. The other two panels of the same figure show the analogous outcome on damage and regenerating tissue.

In conclusion, although simple, and also considering the limited experimental data, the present model suggests theoretical points of intervention, although the way to achieve the predicted effects through therapeutics may be difficult. This difficulty arises in part because our model does not include additional details such as the relevant signaling pathways, for

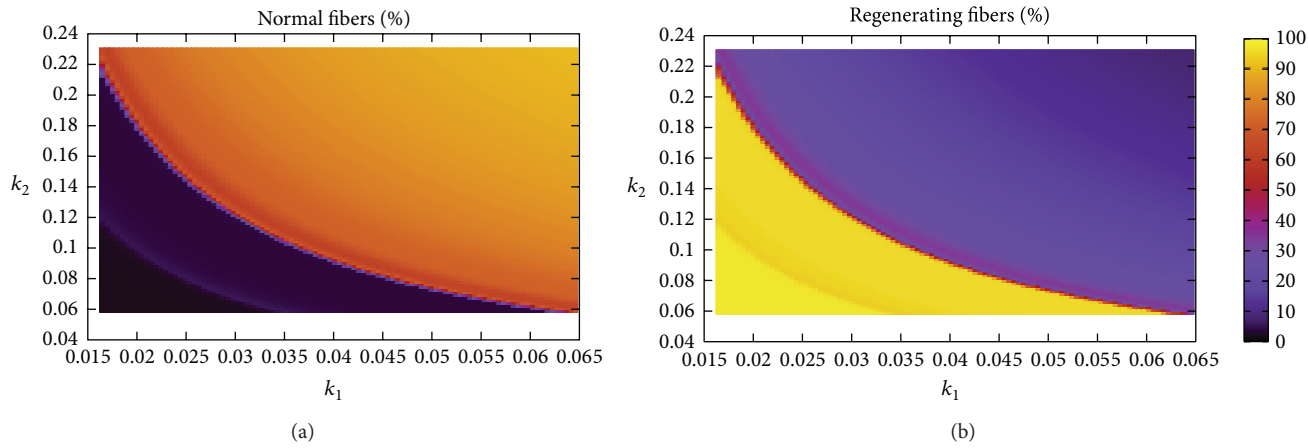


FIGURE 8: The effect of varying the parameters k_1 and k_2, while fixing all other parameters, on the asymptotic (at week 50) percentages of normal and regenerating fibers. These figures indicate that varying both parameters has a stronger effect than either one of them alone, as indicated by the curve between the dark and bright regions.

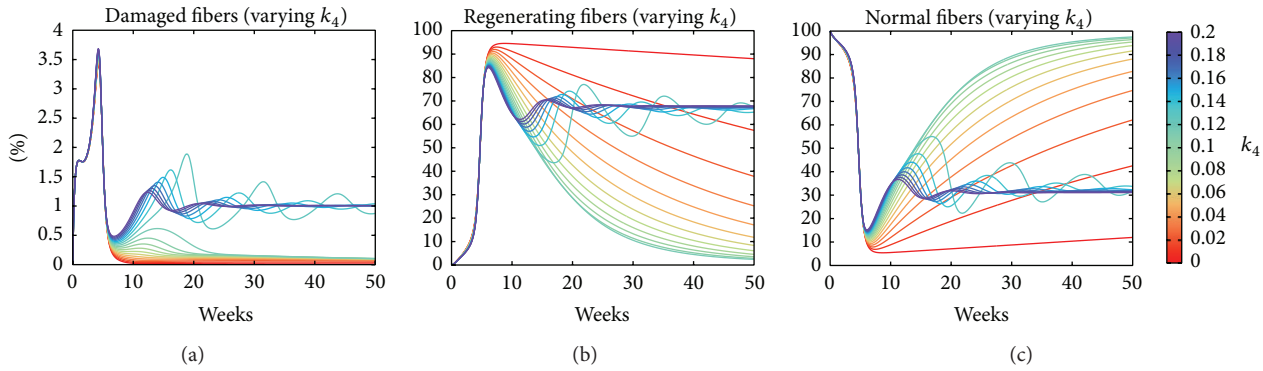


FIGURE 9: The effect of varying the parameter k_4, while fixing all other parameters, on the trajectories of fibers. As the value of k_4 increases, the percentage of damaged fibers decreases (panel (a)) while the percentage of normal (c) fiber increases and eventually reaches 100% (disease-free state), when k_4 is optimal. Beyond that, however, an oscillatory behaviour appears, where the damage is persistent and most of the tissue is regenerating (b).

example, NF-kB, the changes in the population of M1 and M2 macrophages, nor does it keep track of the many potential cytokines and chemokines that are involved.

Nevertheless, the discussion above suggests potential treatment approaches. For example, a treatment approach would decrease antigen presentation by macrophages, thus reducing T-helper activation and consequent cytotoxic stimulation (parameters k_1 and k_2); a second approach attenuates the destruction of normal tissue by cytotoxic T-cells (parameter k_5), while a third approach would drive the generation rate of healthy fibers k_4 to near optimal. In practical terms, however, the suggested therapeutic approaches may not be effective because the model lacks many additional and critical details as noted above. It would be interesting to investigate whether a modification of the model to account for the aforementioned deficits would still identify limiting antigen presentation or cytotoxicity as the main potential candidates for mitigating the immune system effects on the dystrophic pathophysiology.

Conflict of Interests

The authors declare that there is no conflict of interests.

Authors' Contribution

Abdul Salam Jarrah and Filippo Castiglione contributed equally to this work.

Acknowledgments

The authors thank the referee for the valuable comments and suggestions. Abdul Salam Jarrah thanks Stafan Hoops for his help with the software COPASI. Abdul Salam Jarrah was partially supported by a faculty research grant (FRG3). Filippo Castiglione thanks the European Commission for partially funding this study under the 7th Framework Programme (MISSION-T2D project no. 600803).

References

[1] D. J. Blake, A. Weir, S. E. Newey, and K. E. Davies, "Function and genetics of dystrophin and dystrophin-related proteins in muscle," *Physiological Reviews*, vol. 82, no. 2, pp. 291–329, 2002.

[2] J. G. Tidball and M. Wehling-Henricks, "Evolving therapeutic strategies for Duchenne muscular dystrophy: targeting downstream events," *Pediatric Research*, vol. 56, no. 6, pp. 831–841, 2004.

[3] A. E. H. Emery, "The muscular dystrophies," *The Lancet*, vol. 359, no. 9307, pp. 687–695, 2002.

[4] G. Bulfield, W. G. Siller, P. A. L. Wight, and K. J. Moore, "X chromosome-linked muscular dystrophy (mdx) in the mouse," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 81, no. 4, pp. 1189–1192, 1984.

[5] M. Durbeej and K. P. Campbell, "Muscular dystrophies involving the dystrophin-glycoprotein complex: an overview of current mouse models," *Current Opinion in Genetics and Development*, vol. 12, no. 3, pp. 349–361, 2002.

[6] C. A. Collins and J. E. Morgan, "Duchenne's muscular dystrophy: animal models used to investigate pathogenesis and develop therapeutic strategies," *International Journal of Experimental Pathology*, vol. 84, no. 4, pp. 165–172, 2003.

[7] T. A. Partridge, "The mdx mouse model as a surrogate for Duchenne muscular dystrophy," *The FEBS Journal*, vol. 280, no. 17, pp. 4177–4186, 2013.

[8] J. D. Porter, W. Guo, A. P. Merriam et al., "Persistent overexpression of specific CC class chemokines correlates with macrophage and T-cell recruitment in mdx skeletal muscle," *Neuromuscular Disorders*, vol. 13, no. 3, pp. 223–235, 2003.

[9] M. J. Spencer, C. M. Walsh, K. A. Dorshkind, E. M. Rodriguez, and J. G. Tidball, "Myonuclear apoptosis in dystrophic mdx muscle occurs by perforin- mediated cytotoxicity," *The Journal of Clinical Investigation*, vol. 99, no. 11, pp. 2745–2751, 1997.

[10] M. J. Spencer and J. G. Tidball, "Do immune cells promote the pathology of dystrophin-deficient myopathies?" *Neuromuscular Disorders*, vol. 11, no. 6-7, pp. 556–564, 2001.

[11] M. Wehling, M. J. Spencer, and J. G. Tidball, "A nitric oxide synthase transgene ameliorates muscular dystrophy in mdx mice," *Journal of Cell Biology*, vol. 155, no. 1, pp. 123–132, 2001.

[12] J. D. Porter, S. Khanna, H. J. Kaminski et al., "A chronic inflammatory response dominates the skeletal muscle molecular signature in dystrophin-deficient mdx mice," *Human Molecular Genetics*, vol. 11, no. 3, pp. 263–272, 2002.

[13] N. P. Evans, S. A. Misyak, J. L. Robertson, J. Bassaganya-Riera, and R. W. Grange, "Immune-mediated mechanisms potentially regulate the disease time-course of duchenne muscular dystrophy and provide targets for therapeutic intervention," *PM&R*, vol. 1, no. 8, pp. 755–768, 2009.

[14] N. P. Evans, J. A. Call, J. Bassaganya-Riera, J. L. Robertson, and R. W. Grange, "Green tea extract decreases muscle pathology and NF-κB immunostaining in regenerating muscle fibers of mdx mice," *Clinical Nutrition*, vol. 29, no. 3, pp. 391–398, 2010.

[15] M. J. Spencer, E. Montecino-Rodriguez, K. Dorshkind, and J. G. Tidball, "Helper ($CD4^+$) and cytotoxic ($CD8^+$) T cells promote the pathology of dystrophin-deficient muscle," *Clinical Immunology*, vol. 98, no. 2, pp. 235–243, 2001.

[16] A. N. Kolmogorov, "Uber das logarithmisch normale Verteilungsgesetz der Dimensionen der Teilchen bei Zerstuckelung," *Doklady Akademii Nauk*, vol. 31, pp. 99–101, 1941.

[17] S. Hoops, R. Gauges, C. Lee et al., "COPASI—a COmplex PAthway SImulator," *Bioinformatics*, vol. 22, no. 24, pp. 3067–3074, 2006.

[18] Y. Tanabe, K. Esaki, and T. Nomura, "Skeletal muscle pathology in X chromosome-linked muscular dystrophy (mdx) mouse," *Acta Neuropathologica*, vol. 69, no. 1-2, pp. 91–95, 1986.

[19] J. X. DiMario, A. Uzman, and R. C. Strohman, "Fiber regeneration is not persistent in dystrophic (mdx) mouse skeletal muscle," *Developmental Biology*, vol. 148, no. 1, pp. 314–321, 1991.

[20] T. Yokota, Q.-L. Lu, J. E. Morgan et al., "Expansion of revertant fibers in dystrophic mdx muscles reflects activity of muscle precursor cells and serves as an index of muscle regeneration," *Journal of Cell Science*, vol. 119, no. 13, pp. 2679–2687, 2006.

[21] J. Stark, C. Chan, and A. J. T. George, "Oscillations in the immune system," *Immunological Reviews*, vol. 216, no. 1, pp. 213–231, 2007.

[22] T. J. Hawke, A. P. Meeson, N. Jiang et al., "p21 is essential for normal myogenic progenitor cell function in regenerating skeletal muscle," *American Journal of Physiology—Cell Physiology*, vol. 285, no. 5, pp. C1019–C1027, 2003.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 264519, 12 pages
http://dx.doi.org/10.1155/2014/264519

Research Article

A Bioinformatics Pipeline for the Analyses of Viral Escape Dynamics and Host Immune Responses during an Infection

Preston Leung, Rowena Bull, Andrew Lloyd, and Fabio Luciani

Inflammation and Infection Research Centre, School of Medical Sciences, The University of New South Wales, Sydney, NSW 2052, Australia

Correspondence should be addressed to Fabio Luciani; luciani@unsw.edu.au

Received 31 March 2014; Accepted 8 May 2014; Published 10 June 2014

Academic Editor: Filippo Castiglione

Rapidly mutating viruses, such as hepatitis C virus (HCV) and HIV, have adopted evolutionary strategies that allow escape from the host immune response via genomic mutations. Recent advances in high-throughput sequencing are reshaping the field of immunovirology of viral infections, as these allow fast and cheap generation of genomic data. However, due to the large volumes of data generated, a thorough understanding of the biological and immunological significance of such information is often difficult. This paper proposes a pipeline that allows visualization and statistical analysis of viral mutations that are associated with immune escape. Taking next generation sequencing data from longitudinal analysis of HCV viral genomes during a single HCV infection, along with antigen specific T-cell responses detected from the same subject, we demonstrate the applicability of these tools in the context of primary HCV infection. We provide a statistical and visual explanation of the relationship between cooccurring mutations on the viral genome and the parallel adaptive immune response against HCV.

1. Introduction

The complete lack or limited efficacy of vaccines for rapidly mutating viruses causing chronic infections (e.g., HIV and HCV) or seasonal pandemics (Influenza) is due to the extreme and rapid adaptation dynamics of these viruses at both the within and between host level. The extremely high mutation rate of these genomes results in single nucleotide polymorphisms (SNPs) emerging in the viral genome in a short time scale, which in turn gives rise to immune escape phenotypes. These mutations result in new viral variants that avoid detection from the adaptive immune responses (T-cell and B-cell responses), which previously targeted the original virus. However, these mutations are likely to have fitness costs and limit the successful transmission of viral escape variants to a new host. Viruses have to compensate for these costs with additional mutations. Therefore, to understand viral immune escape it is important to study the cooccurrence of multiple mutations on the same genome, as these are the source of compensatory mutations that can drive successful transmission of immune escape variants at the population level. An important lesson learned from the failed T-cell based HIV vaccine (STEP) [1] trials is that T-cell induced responses are not protective against the virus if their targets are viral antigens with a high likelihood of beneficial compensatory mutations, which would allow rapid and successful immune escape. Rather, the data suggests that successful vaccines need to induce strong immune responses against epitopes for which viral escape variants are unlikely to establish a successful new viral population. For instance, T-cell responses should be induced on viral epitopes that are associated with significant deleterious effects on virus viability.

1.1. Next Generation Sequencing Technology (NGS). Recent advances in high-throughput sequencing allow researchers to generate very large data sets of pathogen genomes as well as that of host genomes and transcriptomes. These data, although useful, also carry a complexity that requires computational analyses to understand and represent the biological phenomena. NGS, in particular, has become a powerful tool for deep sequencing analyses of highly variable genomic populations, such as those arising during an infection with rapidly mutating pathogens, or metagenomics [2]. Similarly,

in immunology, NGS is increasingly utilized for sequencing of highly polymorphic protein encoding regions, which are key elements of pathogen recognition, including HLA genes, T-cell receptor [3–5], and the B cell receptors [6]. NGS technology has been also utilized to study the evolution of complex viral populations evolving during an infection within host, such as the detection of rare viral variants in hepatitis C virus (HCV) [7], or in HIV [8]. These data have a strong significance in the study of the dynamics of immune escape as it provides deep insights into the kinetics and extent of viral immune escape [9].

1.2. HCV and Host Immune Response. HCV establishes chronic infections in 60–80% of the infected population [10]. Those individuals who have chronic infection outcomes can eventually undergo liver damage and the development of hepatocellular carcinoma (HCC) [11], the fifth most common cancer worldwide [12]. HCV is a single stranded RNA virus; it evolves by rapidly mutating its genome during a single infection with an estimated mutation rate of 1.2×10^{-4} per site per cell [13]. Similar to other RNA viruses, HCV mutates frequently across the genome, resulting in a high degree of heterogeneity between the seven different HCV genotypes, which carry only ~65% identity. The high mutation rate occurs because the viral encoded RNA-dependent RNA polymerase lacks proof-reading capacity, resulting in at least one error in each genome copied [14]. Consequently, HCV exists as a diverse and evolving population within each infected host. However, it is estimated that about a third of the mutations introduced are deleterious [13], hence many new variants are eliminated in a potent negative selection process [15]. These high error rates are also advantageous to the virus, as they drive rapid adaptation to changing environmental landscapes, such as transmission to a new host or an emerging immune escape variant [16, 17]. Therefore, there is a tradeoff between beneficial and deleterious mutants. This rapid evolution is critical for the survival of these viruses during the establishment and maintenance of chronic viral infections. It facilitates viral escape from host immune responses [18] and optimizes replication efficiency.

Whether infected individuals clear HCV or have viral persistence can be determined by the host immune response [9, 19]. In particular, observation of HCV cytotoxic T-cells (expressing CD8+ marker, CD8+ T-cells henceforth) has shown that these cells contribute to the infection's outcome, where virus-specific CD8+ T-cells are crucial in controlling HCV replication [18, 19]. Although HCV specific CD8+ T-cells are found in acute HCV infection, their efficacy in persistent infection is limited by several factors including T-cell exhaustion and immune escape. The drive behind immune escape is due to the error prone replication mechanism of HCV, thus allowing adaptation to the immune selection pressure through genomic sequence mutations [18]. For example, nonsynonymous mutations in the viral genomic sequence can lead to a peptide change at the protein level of the genome. This mutation can abrogate the immune recognition pathway at several levels. For instance, this mutation can impair antigen processing from viral proteins, as well as the presentation of CD8+ T-cell epitope [20].

1.3. Transmitted/Founder Virus. The transmitted/founder virus is the strain that successfully infects a new host after a transmission event. In HCV studies it has been shown that very few (1–3) transmitted/founder viruses are present in acute infections [7]. An additional study has shown that a genetic bottleneck (an event where genetic variation is greatly reduced) occurs later in infection when selective pressure from the host immune response acts against the transmitted/founder virus [7]. The study observed that as a result of genetic bottleneck events, new viral populations emerge in subjects that become chronically infected with HCV. These arising variants were characterized by fixation events, namely, mutations occurring in >90% of circulating viral genomes. Close examination on viral sequences found only a minority of these fixations events were likely to be under immune-driven selection.

1.4. Cooccurrence of Genomic Mutations and Its Impact on the Dynamics of Immune Escape. Most genomic mutations in RNA viruses are deleterious [13] and can lead to the extinction of the viral variant through negative selection. A smaller proportion of these mutations may have beneficial effects on the survival of the virus. For instance, cooccurring mutations where a new mutation can compensate for the fitness loss given by previously generated mutations, termed compensatory mutations, are beneficial to the survival of the virus [21]. Thus cooccurring mutations can lead the virus to immune escape or even drug resistance phenotypes. As a consequence, it is crucial to study the interactions between immune responses and the resulting mutations to better understand the mechanism behind viral escape.

1.5. Visualizing Genomic and Immunological Data. Given the level of complexity involved with HCV and the increasingly vast amount of generated data from technologies like NGS, representation of the virus's genomic information is becoming increasingly difficult. As a result, computational methods are becoming more and more crucial in terms of data processing, analysis, visualization, and ultimately understanding biological information. The need for software packages allowing combined analysis of viral evolution, detection of compensatory mutations, and identification of immune escape patterns is evident. Thus the goal of this study was to develop a set of computational and statistical tools to analyze immune escape viral variants. We address this issue by taking the dynamic HCV genome as an example and applying the viral sequence data to our bioinformatics pipeline.

2. Materials and Methods

2.1. Data

2.1.1. NGS Viral Sequences. Longitudinal viraemic samples were collected and sequenced from a subject (Ch_240) that developed chronic HCV infection from genotype 3a. The longitudinal samples were deep sequenced using NGS 454

Roche (for more details see Bull et al. [7]). Four viraemic samples from Ch_240 were deep sequenced, corresponding to 44 days after infection (DPI), 57 DPI, 220 DPI, and 538 DPI, respectively. 454 Roche data were retrieved in a pair of fasta (*.fna*) and quality score (*.qual*) format file from each individual time point.

2.1.2. CD8+ T-Cell Responses. HCV specific CD8+ T-cell response data were available for Ch_240. These responses were measures of HCV specific CD8+ T-cell responses using ELISPOT assay. This assay detected active T-cell responses against specific epitopes, a short amino acid sequence recognized by CD8+ T-cells via presentation of HLA-epitope complex on the surface of infected cells. Two CD8+ T-cell responses specific for epitopes $_{1602}$RAQAPPPSW$_{1610}$ and $_{1633}$RLGPVQNEV$_{1641}$ were utilized in this study. These epitopes are both located in the NS3 region of the HCV genome (nucleotide region 4000–5499 and amino acid region 1200–1800, according to the HCV gt1 reference genome H77, GenBank accession number AF009606).

2.1.3. Single Genome Amplification (SGA) Sequences. Longitudinal sequence data were retrieved from publicly available data [22]. We analyzed sequences from three time points within the first 4 weeks of acute infection from one subject (10012) infected with HCV genotype 1a.

2.2. Tools for Computational Analysis

2.2.1. Quality Control and Sequence. Each fasta and quality score file are processed by *choplqb.py* with options *−w 8 −t 15* for initial data cleaning. Refined outputs from *choplqb.py* are then given to qualfa2fq.pl to combine the fasta file and quality score file into a single fastq file. This is then used for sequence alignment using the Burrow Wheel algorithm, implemented in the software package BWA [23]. BWA and qualfa2fq.pl are both available at http://sourceforge.net/projects/bio-bwa/files/ (see *README.md* at the GitHub repository located below for more information).

2.2.2. HCV NS3 Reconstruction. Refined NGS data are processed by SamTools [24] (version 0.1.19, http://sourceforge.net/projects/samtools/files/latest/download) to convert the sam file into bam format in order to apply the data to the genome reconstruction software ShoRAH [25] (Short Reads into Assembly Haplotypes). This software allows for reconstruction of partial or full genome sequences from a mixed population of genomes, henceforth reconstruction of viral haplotypes (this software is available at http://www.bsse.ethz.ch/cbg/software/shorah). The options for ShoRAH include *−a 0.05 −w 300* and a length restriction of 1500 nucleotides in the NS3 region 4000–5499. Sets of reconstructed and aligned viral sequence files in fasta format are retrieved using this software. For this analysis, only full-length (1500 nt) reconstructed fasta files (*.popl*) from ShoRAH that carry viral variants with frequency of occurrence of 1% or greater are applied to this pipeline.

2.2.3. Reducing Insertion and Deletion Errors. Viral sequences reconstructedfor each time point are piped into *indelRemover.py* using option *−fq 2* for insertion and deletion reduction. The output is another fasta format file with reduced insertion and deletions around homopolymers regions in the viral sequences.

2.2.4. Analysis of Cooccurring Mutations. Fasta files of viral genome populations are given to *a-smupfi.py* with options *−g −gf 0.01 −e −sc 242 −m 1-2 −s 4000–5499,* which produced four text files containing the combinations of shared mutations and their frequency of occurrence between viral sequences observed in the data set. From these four files, the file with suffix *EasyOutputShared.txt* is used as input data for *javssim.py* with option *−e*. The result provided by *javssim.py* contains a list of shared combinations of mutations. This tool utilizes the Jaccard similarity coefficient, calculated for a given pair of mutations A and B, as $J = N_{AB}/(N_{AB} + N_{A0} + N_{0B})$, where N_{AB} represents the number of variants carrying mutations A and B, N_{A0} represents the number of variants containing A but not B, and N_{0B} represents the number of variants containing B but not A. The algorithm implemented here for the detection of statistically significant pairs was taken from Rhee et al. [26]. Briefly, statistically significant pairs are identified from the expected Jaccard similarity coefficient and its standard error assuming the two mutations independently distributed. J_{RAND} was calculated as the mean Jaccard similarity coefficient after 2,000 random rearrangements of the X or Y vector (containing 0 or 1 for presence or absence of a mutation, resp.). J_{SE} was calculated using a jackknifed procedure, which removed one sequence at a time, repeatedly for each sequence. The standardized score Z, $Z = (J - J_{\mathrm{RAND}})/J_{\mathrm{SE}}$, indicates a significant positive association ($Z > 1.65$) or a significant negative association ($Z < -1.65$) at an unadjusted $P < 0.05$.

The results from this analysis are then found in the file with suffix *EasyOutput.txt* from *javssim.py*. This file without a header can be parsed into *jaccardToCircos.py* with option *−li* for format conversion into a style that can be understood by Circos [27] (software is available at http://circos.ca/) to draw connections between significant cooccurring mutations.

2.2.5. Analysis of Covariance. The covariance analysis was performed on viral sequences translated into amino acid using ExPASy tools [28]. These are then inputted into *omes.py* for calculation of the covariance score, according to the observed minus expected squared (OMES) method [29] as

$$\mathrm{Cov}_{\mathrm{score}} = \frac{\sum_1^L \left(N_{\mathrm{obs}} - N_{\mathrm{exp}}\right)^2}{N_{\mathrm{sequences}}},$$
$$N_{\mathrm{exp}} = \frac{\left(F_{aj} \times F_{bk}\right)}{N_{\mathrm{sequences}}}, \quad (1)$$

where each paired position j and k has residue a and residue b, respectively. F_{aj} is the frequency of occurrence of a at position j and F_{bk} is the frequency of occurrence of b at position k. $N_{\mathrm{sequences}}$ is the number of sequences at positions

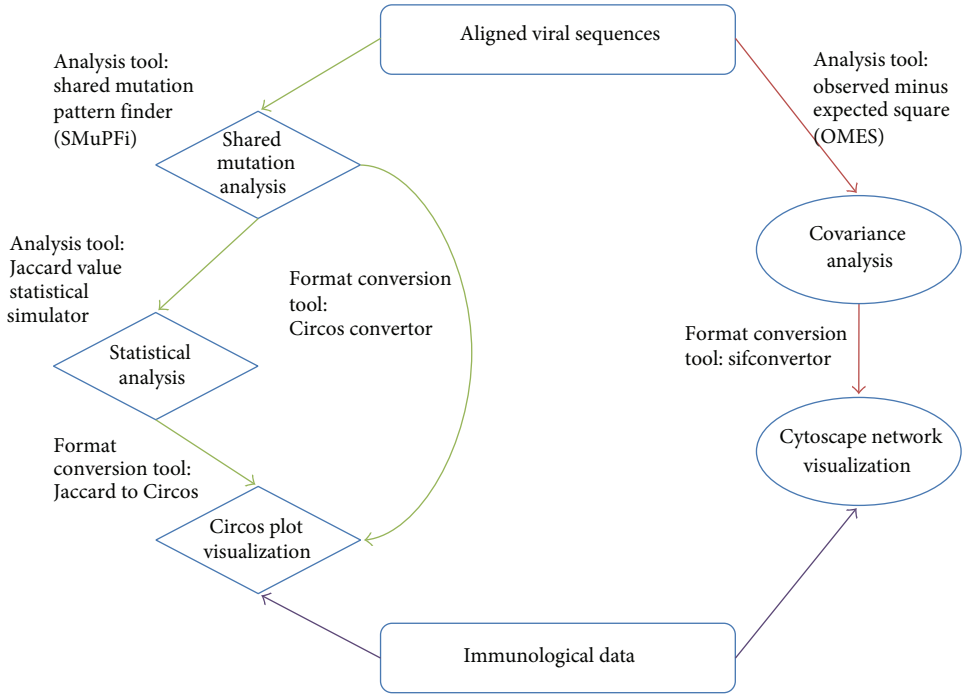


Figure 1: Flowchart representing the bioinformatics pipeline. Shared Mutation Analysis workflow is indicated by green arrows. Covariance Analysis workflow is indicated by red arrows. Both branches of the pipeline require an aligned viral sequence file in fasta format. Immunological data are optional and can be both experimental or bioinformatics predicted.

j and k. N_{obs} is the frequency of occurrence of the pair a and b at positions j and k. L is the number of unique pairs counted at positions j and k. This output is parsed into *sifconvertor.py* to generate a network file (*.sif*), which can be parsed into Cytoscape [30] (software is available at http://www.cytoscape.org/) for simple network visualization. For this analysis only mutation pairs that have a covariance score equal to or greater than 0.5 (cut-off value taken from Aurora et al. [31]) were considered for further analysis.

2.2.6. Histogram Generation. Using the same set of output from *a-smupfi.py*, the file with suffix *EasyOutputUnique.txt* is applied as input into *circosconverter.py* with options *–hi* to generate Circos format histograms. The bam file from the NGS data (also used for generating haplotypes with ShoRAH) is used to identify SNPs within the viral population. This SNP detection was performed using LoFreq (version 0.5.0) [32] (software is available at http://sourceforge.net/projects/lofreq/files/) using default options for additional SNP calling. However, other software can be utilized for SNP detection from NGS bam file. SNPs data are then parsed into *snpExtractor.py* and then those outputs are given to *snpExtractConvertor.py* for Circos format conversion into histogram files.

2.2.7. Phylogenetic Tree Generation. Viral sequences of all four time points from Ch_240 have been appended into one sequence file and given as input to PhyML Ver. 3.0.1 [33] (software available at http://www.atgc-montpellier.fr/phyml/) changing options *S* (Tree topology search operations) to *Best of NNI and SPR (extensive tree search)* and *B* (nonparametric bootstrap analysis) to *1000*. The phylogenetic tree was generated under the substitution model HKY85 using estimated γ distribution on sites. The output file from PhyML is represented and visually edited through the software FigTree Ver. 1.3.1 (software available at http://tree.bio.ed.ac.uk/software/figtree/).

Script names in italics are original tools and are available at the GitHub repository: https://github.com/PrestonLeung/SMuPFi-Repository.

3. Results

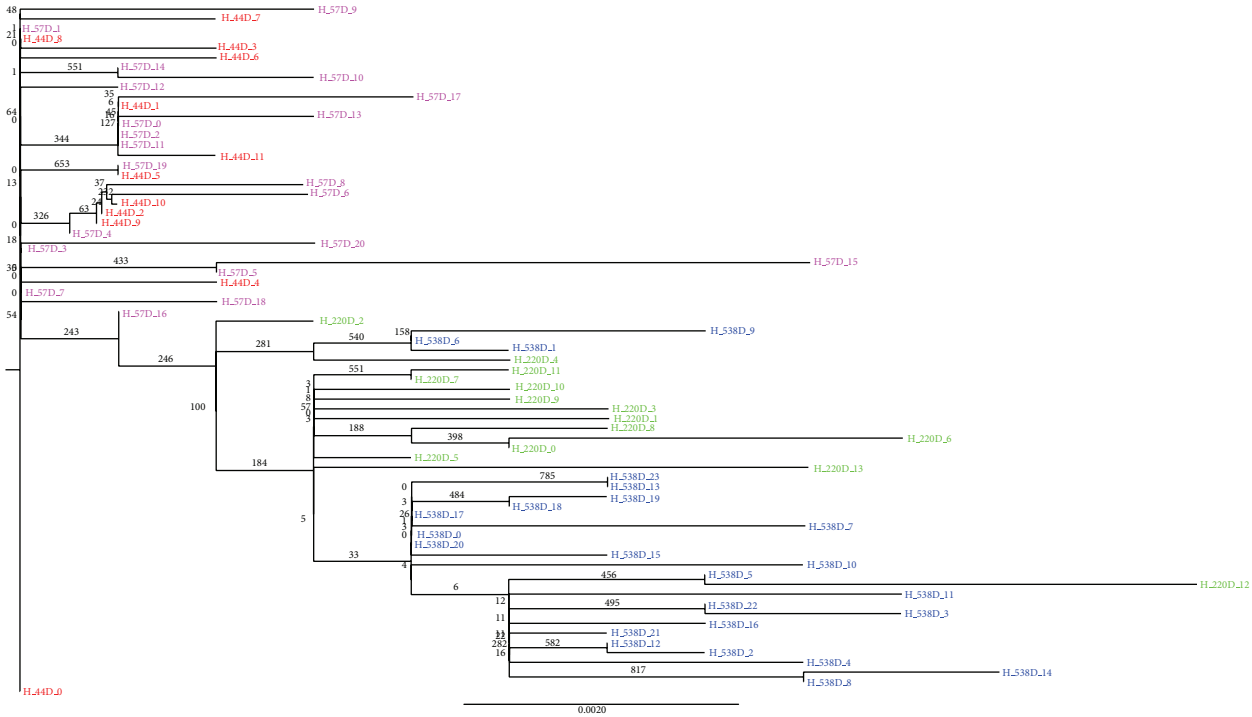
3.1. An Overview of the Pipeline. A workflow of the proposed pipeline for the analysis of genomic and immunological data related to immune escape during HCV infection is shown in Figure 1. The pipeline builds a path that connects NGS data, viral sequence quality control, alignment software, haplotype reconstruction (e.g., using ShoRAH), and identification of the distribution of variants sharing a number of SNPs. It also provides graphical representation of these findings through Circos and Cytoscape. In this pipeline, we present the Shared Mutation Pattern Finder (SMuPFi), a novel algorithm package that analyses cooccurring mutation patterns with a series of supporting tools inside the package. Figure 1 shows the workflow that describes the cooccurring mutation analysis and the covariance network analysis. The pipeline uses viral sequences and immunological data as the input source where the region of interest is selected based on immunological data. The cooccurring mutation analysis uses viral sequences as the main data source. Combinations

of mutation with length of 2 and occurring on the same viral genome are then selected. Combinations of mutation pairs that are shared by more than two viral variants are identified for further analysis. This dataset is then used to identify those combinations of coupled mutations that occur in the viral population at a frequency higher than that expected by random mutation. These are recognized through a test for statistical significance implementing the Jaccard similarity coefficient (see Section 2). Statistically significant cooccurring mutations are identified from the expected distribution of cooccurrence assuming independency of the two mutations of interest. These pairs are then graphically represented via Circos plots. The covariance network analysis similarly uses viral sequences as input data. It identifies pairs of mutated sites and utilizes the paired positions and frequencies to calculate a covariance score. This score reveals those pairs that are likely to have a relationship such that their cooccurrence is more than an expected value, which is calculated as the product of frequencies of each residue in the pair, divided by the total number of sequences (OMES see Section 2). The data is then represented in a network using Cytoscape.

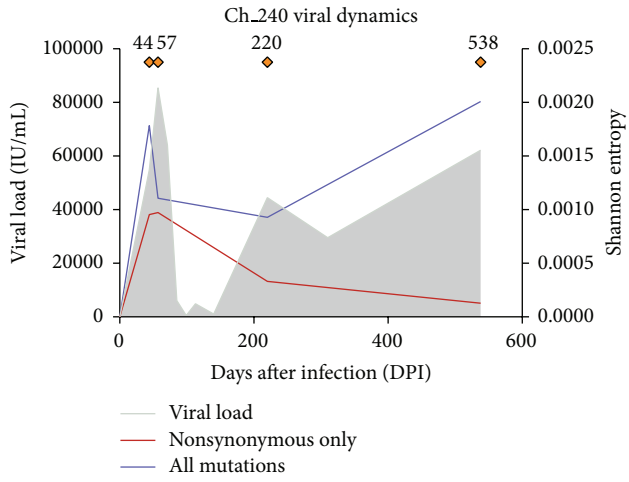
3.2. Application of the Pipeline to HCV. Virological and immunological data from a subject (Ch_240) chronically infected with HCV has been applied to the pipeline using deep sequencing data available from four time points (44 DPI, 57 DPI, 220 DPI, and 538 DPI). The input data consisted of viral sequences from a segment of the NS3 region (nucleotide region 4000–5499 and amino acid region 1200–1800, see Section 2 for more details) and immunological data, which were available ELISPOT measurements of CD8+ T-cell responses specific for two HCV antigenic epitopes $_{1633}$RLGPVQNEV$_{1641}$ and $_{1602}$RAQAPPPSW$_{1610}$. Both of these epitopes were located within the sequenced NS3 region.

Figure 2 shows a phylogenetic tree representation of the viral sequences from the partial NS3 region of the HCV genome of subject Ch_240 displaying the genetic distances measured by nucleotide differences using the transmitted/founder virus (labeled H_44 DPI_0) as the root of the tree. In this subject it is evident that the viral population significantly evolved from the transmitted/founder virus during the early acute phase of the infection with rapid genomic diversification undergoing sequential genetic bottlenecks events. In this figure, the viral dynamics plot (zoomed panel in Figure 2) describes the level of variability in the genome using Shannon entropy measured across the full HCV genome. This plot shows the first genetic bottleneck occurring at approximately 100 DPI where viral diversity significantly decreases, along with the viral load and the loss of the dominant viral variants [7]. At 200 DPI, the viral load increases, along with the number of distinct circulating NS3 variants, thus indicating a new viral population characterizing the chronic phase of infection. Of note, a second genetic bottleneck occurs between 220 and 538 DPI. While the figure shows sequences from 44 DPI and 57 DPI evolving into the latter group (220 DPI and then 538 DPI), notably there is a large time period of approximately 300 days between 220 DPI and 538 DPI. This rapid evolution is a major component driving the success of the virus to establish chronic infection and escape the host immune response. These mutations are further explored in the following analyses.

3.3. Identification of Cooccurring Genomic Mutations. We considered the evolutionary dynamics of circulating viral genomes and HCV specific CD8+ T-cell responses over the course of an infection. Figure 3 shows the evolutionary dynamics of viral genomes in subject Ch_240 and the distribution of cooccurring genomic mutations and their relationship with viral diversity and T-cell responses. Figure 3(a) shows the viral dynamics as already explained in Figure 2. Figure 3(b) shows the experimental results (ELISPOT data) detailing the measurement of two HCV specific CD8+ T-cell responses targeting epitopes $_{1602}$RAQAPPPSW$_{1610}$ (green) and $_{1633}$RLGPVQNEV$_{1641}$ (pink) within the NS3 region. Immune responses are first detected at 44 DPI but at very low amount. In this context of early onset of CD8+ T-cell responses, and a concomitant genetic bottleneck event characterizing the circulating HCV viral population, we investigated the hypothesis that these immune responses were driving evolution of HCV genomes. We therefore assessed the distribution of mutations and their cooccurrence before and after this genetic bottleneck, as well as the evolution of these mutants in relation to the appearance of the two CD8+ T-cell responses identified within this region. Viral diversification was very limited within the first 57 DPI, and no statistically significant cooccurring mutations were observed. In particular, nonsynonymous substitution P1606S was observed within the CD8+ T-cell epitope $_{1602}$RAQAPPPSW$_{1610}$ at low frequency at 44 DPI but this mutation was not detected at 57 DPI, most possibly because it was at a frequency of occurrence lower than 1% (the minimum threshold detected after error correction of sequencing data). CD8+ T-cell responses against the two epitopes $_{1633}$RLGPVQNEV$_{1641}$ and $_{1602}$RAQAPPPSW$_{1610}$ were detected at increased magnitude at 85 DPI, close to the time of the first genetic bottleneck (Figure 3(a)). This was followed by the emergence of a new viral population after 100 DPI (Figure 2). Following this genetic bottleneck, a new mutation P1606L was identified within the epitope region $_{1602}$RAQAPPPSW$_{1610}$ (Figure 3(e)), which was shared among 80% of the viral population. This mutation is therefore likely to be an immune escape variant. Another mutation, G1671R, was then identified to cooccur with the immune escape mutation P1606L at 220 DPI. The mutation G1671R was not present when the P1606S was first identified, and these data suggest G1671R is potentially compensating for P1606L within the epitope region targeted by CD8+ T-cell responses. Secondly, P1606S occurred at a very low frequency at 44 DPI followed by nondetection in 57 DPI (perhaps due to extremely low frequency of occurrence), indicating the possibility that it is an individual deleterious mutation, which restricts the fitness or success of the escape variant. This analysis hence indicates that detection of cooccurring mutations over the course of the infection can be utilized to detect key patterns of immune escape.



(a)



(b)

Figure 2: Phylogenetic tree representation of viral sequences obtained from Ch_240. Phylogenetic tree of circulating HCV variants from sequences derived from the NS3 region (nucleotide region 4000–5499, amino acid region 1200–1800). This tree shows the significant rapid evolution of HCV genome over the course of the infection from acute phase (in red viral sequences from 44 DPI, in pink those for 57 DPI) through the chronic phase of infection (in green sequences from 220 DPI, and in blue from 538 DPI). The infection is characterized by two sequential genetic bottlenecks, one soon after the acute phase of infection, the other after 220 DPI, both indicating substantial changes in the circulating viral populations. The label H_44D_0 indicates the transmitted/founder virus used to root the phylogenetic tree. The numbers at branches indicate the bootstrap value after resampling 1000 times. Viral Dynamics plot (b) shows the viral load (grey shading) over time, while the orange diamonds indicate available deep sequence data (from left to right: 44, 57, 220, 538 DPI). Shannon entropy calculated from the full distribution of HCV genomic mutations and the one from the distribution of nonsynonymous mutations only are indicated by a purple line and red line, respectively.

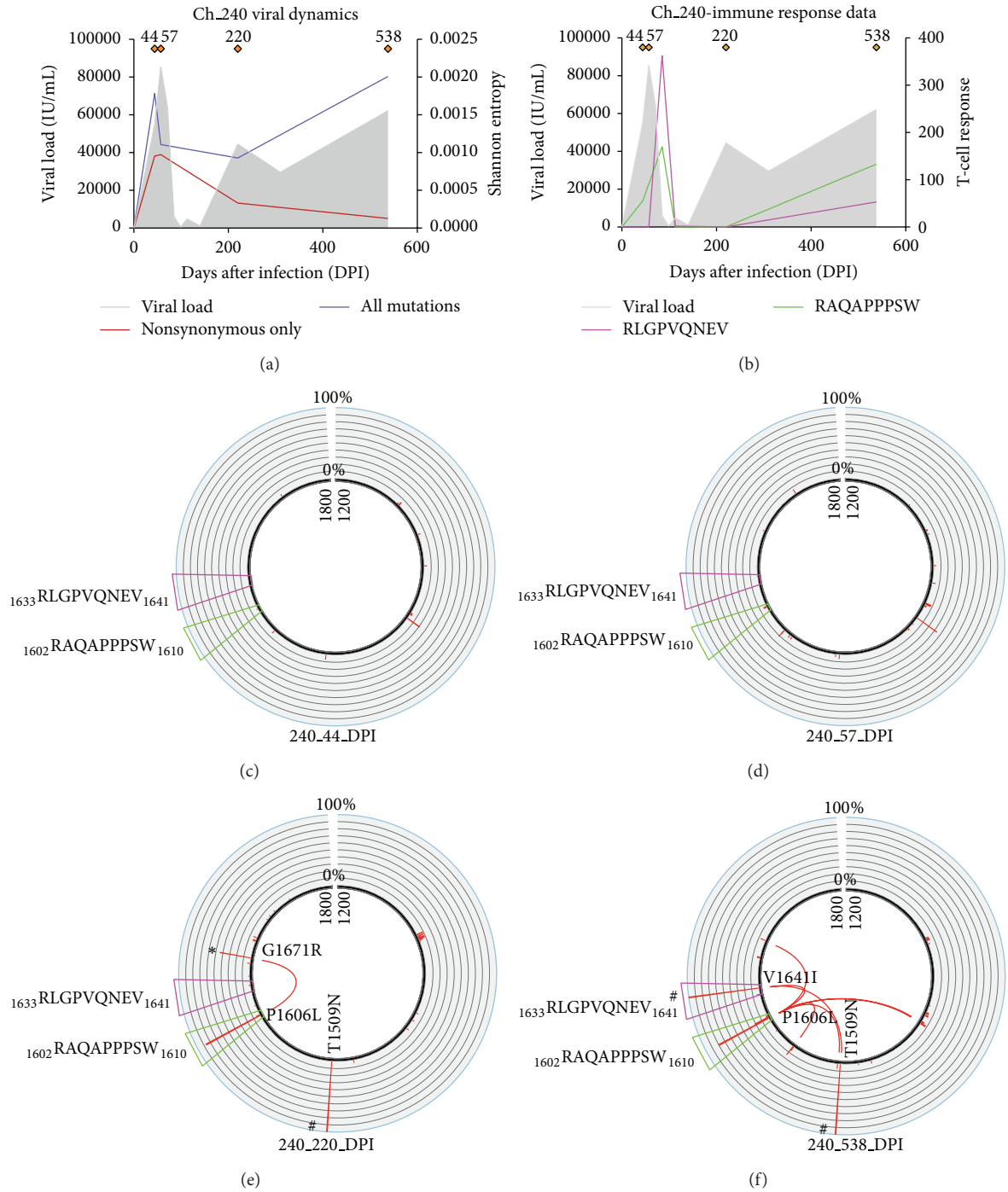


FIGURE 3: A representative example of the output of the pipeline applied on longitudinally collected viral sequences and immune responses data from a patient chronically infected with HCV. (a) The same viral dynamics plot carried over from Figure 2. (b) The experimental results from ELISPOT detailing the measurement of two HCV specific CD8+ T-cell responses targeting epitopes $_{1602}$RAQAPPPSW$_{1610}$ (green) and $_{1633}$RLGPVQNEV$_{1641}$ (pink). Immune responses are first detected at around 50 DPI at low amount. These responses then increase in magnitude at 80–100 DPI and then decline over the course of the infection. (c–f) Circos plots representing the partial NS3 region (amino acid region 1200–1800) in Ch_240 that contain the two epitopes targeted by the CD8+ T-cell responses in (b). The red histogram around the circle shows the frequency of occurrence of nonsynonymous mutations with a scale from 0% to 100% at that site. These two CD8+ T-Cell epitopes $_{1602}$RAQAPPPSW$_{1610}$ (green trapezium) and $_{1633}$RLGPVQNEV$_{1641}$ (pink trapezium) are also represented. Fixation events (mutations at frequency >90%) are represented with # in (e) and (f). The arcs shown in the inner circular area represent statistically significant cooccurring pairs of mutations that are shared by two or more viral variants. The asterisk (∗) in (e) denotes the potential compensatory mutation at position G1671R cooccurring with immune escape variants identified within $_{1602}$RAQAPPPSW$_{1610}$ (green).

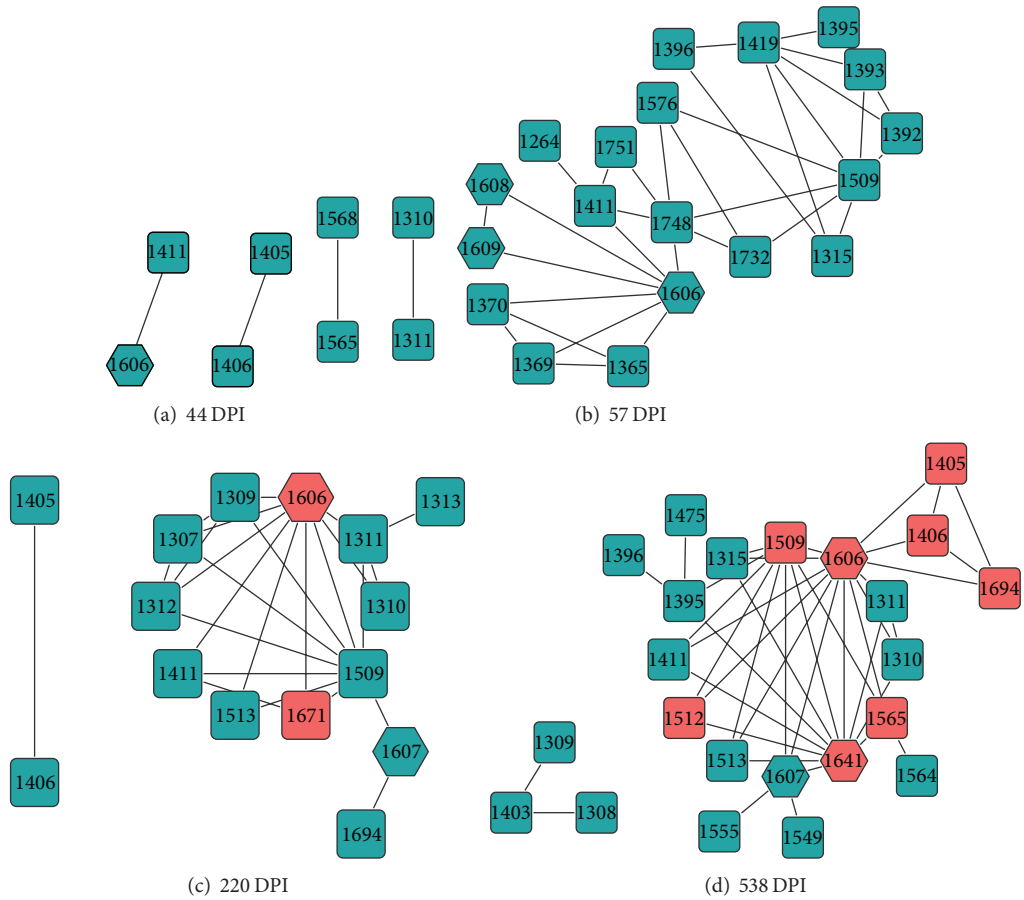


FIGURE 4: A representative example of the covariance network analysis. Network representation pair-wise covariance analysis of nonsynonymous mutations detected over the course of infection within the NS3 region (nucleotide region 4000–5499, amino acid region 1200–1800) in subject Ch_240. Nodes presented in the network are positions in the viral sequence found to have significant covariance values that indicate strong covariation between two mutation sites. In each network, the degree of connectivity is increasing from left to right. Hexagonal nodes denote mutation positions that lie within an epitope region, while those in red are mutations that have been detected as statistically significant cooccurring mutations using the Jaccard similarity coefficient. Covarying mutations, which are not shared between two or more viral variants, have their nodes colored in green.

As infection progressed, increased variability in the NS3 region at 538 DPI was observed (Figure 3(e)) approximately 300 days after the previous sequenced time point, despite the overall decline in viral diversity across the full genome (Figure 3(a)). This information suggests that the new viral population is evolving in a specific direction, where HCV viral variants possess mutations that enable the evasion of the CD8+ T-cell responses targeting the NS3 region. As observed in Figure 3(f), the paired mutations P1606L and G1671R that were detected at 220 DPI were lost at 538 DPI. New HCV genomic variants were identified at 538 DPI, with new mutations all cooccurring with P1606L, and a fixation event at V1641I. It is important to note that all HCV genomes at 538 DPI carry the epitope variant $_{1633}$RLGPVQNEI$_{1641}$. Closer examinations revealed that all cooccurring mutations were connected with either position P1606L, V1641I, or both. We also found a fixation event at position T1509N. However this was not a compensatory mutation as the same mutation was already fixed at 220 DPI.

3.4. Covariance Network Analysis. We then analyzed, in a more general fashion, the network of genetic mutations that were identified during the course of the infection in subject Ch_240. To do so, we constructed a covariance score (see Section 2) between all the pairs spanning the entire region (amino acid region 1200–1800). Figure 4 shows the network representing the evolving pattern of the connections between pairs of sites with mutations in the partial NS3 region of subject Ch_240. Each position with a mutation shown on the network is referred to as a node and each line drawn with a neighboring node is referred to as a connection. The most striking feature of the covariance network analysis is the evolving pattern of connections with node 1606. This node, which represents the mutations P1606S (44 DPI) and P1606L (220 DPI and 538 DPI), was observed in the epitope $_{1602}$RAQAPPPSW$_{1610}$ at the earliest time point and was subsequently detected in all other time points, with new connections at every time point.

Covariance network representing 44 DPI (Figure 4(a)) shows a small number of nodes with P1606S connecting to only one neighbor. Diversification of connections is seen at 57 DPI (Figure 4(b)) where the network is more complex due to an increase in the number of nodes and higher degree of connections between nodes. At 220 DPI (Figure 4(c)) an emerging pattern is revealed, where mutation P1606L appears to be a "hub" mutation, having connections with the majority of nodes within the network. 538 DPI (Figure 4(d)) shows a further increase in the number of new connections stemming from P1606L. In particular, the node corresponding to V1641I is another mutation that lies within the CD8+ T-cell epitope $_{1633}$RLGPVQNEV$_{1641}$. Close examination reveals that most nodes detected at 538 DPI (Figure 4(d)) are connected to P1606L or V1641I. These data suggest that the mutants P1606L and V1641I form a potential pair of mutations that is critical for viral escape dynamics.

3.5. Comparison of Cooccurring Mutation Analysis and Covariance Network Analysis. In the analysis of cooccurring pairs of genomic mutation (Figure 3) we showed that HCV immune escape dynamics are characterized by the existence of potential compensatory mutations, which characterize the HCV viral populations, thus suggesting that a strong adaptation mechanism is at play. The covariance network analysis shown in Figure 4 highlights the increasing complexity between the distributions of covarying mutations with the evolution of the infection. These data suggest that cooccurring mutations that affect immune response (hexagonal nodes in Figure 4) are also part of the set of genomic mutations that form the "hub" of the network of covarying mutations (see red nodes in Figure 4). For example, it is interesting to note how the covariance network shows an evolving and complex network of nodes, which supports the appearance of the G1671R being a compensatory mutation to P1606L at 220 DPI after the first genetic bottleneck, when increasing CD8+ T-cell responses were observed. Notably, the appearance of a novel viral population at 538 DPI characterized by the appearance of the pair P1606L and V1641I again reveals a new pattern of connected genetic mutations that contributes to the viral escape dynamics and eventually to viral chronic persistence in subject Ch_240. Interestingly, the network analysis between 44 and 220 DPI showed a rapid modification of the network of covarying mutations, which revealed unstable distribution of mutations connected to the mutation P1606L. This suggests that without the cooccurrence of G1671R at 220 DPI, the immune escape mutation P1606S carries a significant fitness cost, which may limit the survival capacity of this escape variant against CD8+ T-cell response. This is supported by the absence of P1606S at 57 DPI, where there may be a deleterious effect hindering the survivability of the variant.

At 538 DPI (Figure 4(d)), approximately 300 days after the previous time point, the covariance network highlights an increasing complexity of the network, where the majority of the nodes previously observed at 220 DPI disappear and only a few nodes are carried over. In particular, the mutation P1606L is characterized by a completely new set of covarying mutations, again confirming the results from the analysis of cooccurring mutation analysis. This analysis also revealed several highly connected genomic mutations that were also identified from the analysis of cooccurring mutation using the Jaccard similarity coefficient (red nodes in Figure 4). In particular, the fixation at position T1509N is also found as a highly connected node at 538 DPI. Moreover, this analysis identified a triplet of genetic mutations (K1405N, C1406S, and L1694R) that were also connected to the immune escape mutation P1606L, clearly indicating the existence of a subnetwork of evolving genomic variants.

3.6. Pipeline Validation with SGA Data. We validated the proposed bioinformatics pipeline with the analysis of sequences obtained from a subject infected with HCV (Sub 10012, data retrieved from Li et al. [22]) within the first month of infection (measured by weeks after infection or WPI). Given the early stage of infection, we hypothesize the absence of adaptive immune responses targeting those viral populations. Indeed, phylogenetic analysis of sequences from 5′ end to partial NS2 of the HCV genome showed the presence of a random evolution of three major viral populations arising from three transmitted/founder variants that successfully started the infection (Figure 5). The Circos plots show an overall increase in frequency of occurrence of these mutations over time, suggesting the presence of diversifying viral populations without immune pressure. However, despite this apparent random evolution, we identified several cooccurring mutations (Circos plots in Figure 5) that showed highly connected mutation patterns between these three viral populations. Therefore, there is evidence of inheritance of specific mutations throughout the viral evolution, with mutations that occur very early in infection being maintained during the generation of new variants. The covariance networks across the three time points (Figure 5) illustrate the positions where the inherited mutations occur. This highlights several pairs of genomic sites that mutate in each of these populations, thus representing hubs of the evolutionary dynamics of HCV genome during early infection. The comparison of subject 10012 and subject Ch_240 clearly indicates a very different evolutionary pattern driven by the presence of immune response.

4. Discussion

In analyzing viral sequences through the proposed pipeline, this study has revealed a pair of mutations within a region of the viral genome that may form the hub of a network of covarying mutations allowing viral persistence in subject Ch_240. Using the proposed bioinformatics pipeline we addressed the details of immune escape from longitudinal observations of viral evolution of HCV infection and provided insight into the evolution of the virus in relation to the selective pressure exerted by CD8+ T-cell immune responses. Moreover, this analysis provided evidence that the cooccurrence of P1606S and V1641I may be central to the success of immune escape variants against CD8+ T-cell responses targeting HCV during the establishment of chronic

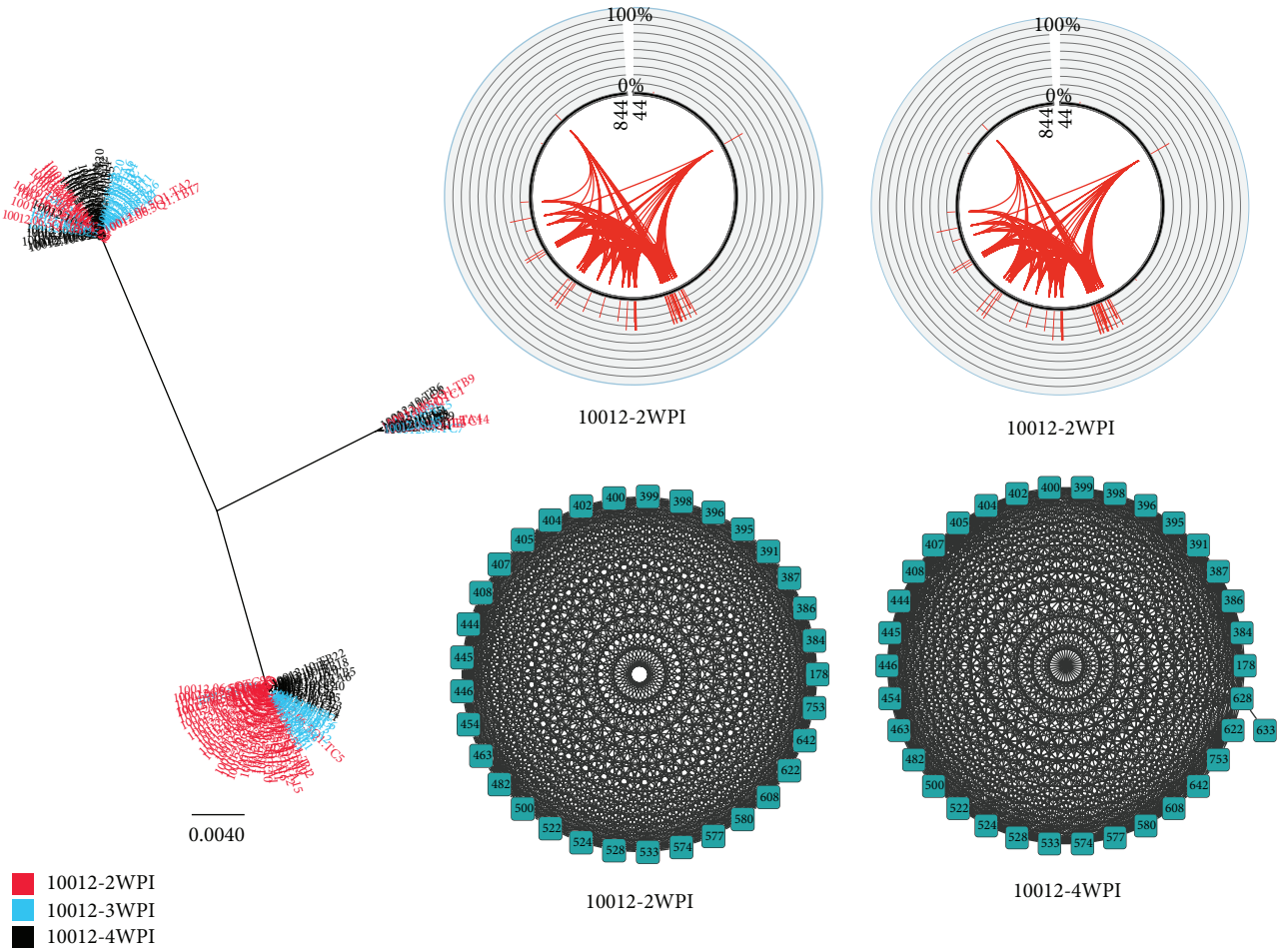


Figure 5: A representative example of a HCV infection presenting highly diverse viral population. Single Genome Amplification (SGA) data from longitudinally collected HCV populations over three time points (2, 3, and 4WPI) taken from a subject (10012) infected with HCV genotype 1a. This analysis was based on sequences from partial Core protein, p7, E1, and E2 protein of HCV from 2WPI and 4WPI (3WPI not shown). Unrooted phylogenetic tree displays three major subpopulations of viruses. A substantial viral diversification is observed since 2 weeks after infection. During this early stage of infection it is unlikely that HCV specific T-cell and B-cell responses are targeting the infection. However, the Circos plots highlight the presence of highly connected variants with specific patterns of cooccurring mutations. This is further validated from the covariance network showing high number of pairs of sites, which are maintained during the early phase of infection.

infection. We showed that HCV evolution under CD8+ T-cell response is characterized by a complex evolving pattern of mutations that consists of mutations in multiple regions functioning as a whole to provide the virus with the ability to escape immune pressure. This analysis, although limited to a small portion of the HCV, provides useful information in identifying potential factors that contributes to the virus's overall escape outcome.

In this study we presented a novel bioinformatics approach for the identification of key viral mutations events that dictate success of viral escape in the establishment of a chronic HCV infection. The output of our computational analysis offers a detailed description of the complex patterns characterizing immune escape dynamics during HCV infection and can therefore be relevant in studying immune escape dynamics in HCV and more in general in other rapidly mutating viruses. In combination with experimental data of CD8+ T-cell responses, our analysis provided a novel method to characterize the dynamics of compensatory mutations. As experimental measures, such as ELISPOT, only test CD8+ T-cell responses against specific epitopes of 8–10 amino acids, analysis of viral genomes is required to explore the distribution of other mutations outside the epitope region, which can serve as compensatory mutations. This analysis has remained elusive because of the lack of appropriate deep sequencing data to measure low frequency variants and the lack of appropriate methods to link distant mutations. With the rise of NGS technologies, and the development of new computational methods for haplotype reconstruction from NGS data [34], this information is becoming accessible. Hence the proposed bioinformatics pipeline is one of the first proposed to provide an exhaustive scenario of immune escape dynamics that takes both genomic and immunological data as input variables.

In the bioinformatics pipeline, we have also provided an array of graphical representation of evolving genomic

mutations connected to each other. With the increasing availability of data and reduction of NGS cost, analyses using computational pipelines are necessary to unravel complexity of large data sets. Moreover, computational analyses can be utilized to minimize experimental costs and reduce time for manually laborious data processing. Since there is no direct monetary penalty in reruns of computational pipelines, exploratory procedures can be done on data with less limitation. For instance the pipeline could be used to obtain preliminary data on viral sequences prior to deciding whether certain regions of the HCV genome are worth spending resources, such as limited samples, or employing assays requiring a large number of cells to achieve sufficient specificity and sensitivity (e.g., ELISPOT or tetramer staining). Bioinformatics predictions of viral epitopes targeted by CD8+ T-cell responses are also available (see http://www.immuneepitope.org). These data can be substituted into experimental measurements and integrated in this pipeline.

The study performed on the partial NS3 region of Ch_240 is only one of many ways of using the pipeline. The tools designed in this pipeline are implemented specifically so that each component of the pipeline can be applied separately. Moreover, different types of viral sequences can be considered as input data. For example, haplotype reconstruction sequences from NGS data can be replaced with single genome amplification data, which are often used to study viral evolution [22] (see Figure 5). Furthermore, supporting tools in the package are able to take in generic fasta sequences and convert these into a format that can be processed in the pipeline.

There are a number of limitations in the proposed pipeline. For example, in this study, only pairs of cooccurring mutations were utilized, in order to limit the complexity of the results. Investigation of a higher number of cooccurring mutations (such as triplets or quadruplets) is possible. However as the length of the viral sequence under investigation increases, the number of cooccurring mutations will also increase (see for instance Figure 5). Using the pipeline without NGS data is also advantageous because current analysis suffers from highly error prone data. For instance, our analysis with haplotypes for subject Ch_240 was based on 454 Roche sequences data, which have been used to reconstruct viral haplotypes with computational demanding software packages. Although the viral genome reconstruction was successful in our study, this method still presents a high false positive rate when reconstructing low frequency variants [35, 36]. However, to address the issue, we have chosen to consider only viral variants reconstructed as haplotypes with a frequency of occurrence greater than 1%.

This computational and statistical framework can also be applied to other viruses and to identify more complex patterns of immune escape or drug resistance. For instance, understanding the dynamics of escape variants against both T and B cell responses. This is a common feature during infections with rapidly mutating viruses, such as HIV [37]. Moreover, the use of Jaccard coefficient allows the identification of specific patterns of mutations that are likely to cooccur more than random expectation. This could be for instance also applied to the detection of drug resistance mutations and for the identification of compensatory sites. The network of mutations performed with the covariance analysis holds a broader goal, and that is to screen viral sequences for major mutating sites or "hubs," identifying sites that are mutating at a significantly high rate across the full genome.

5. Conclusion

This work proposed a novel bioinformatics pipeline for the analysis of immunological and virological data of viral infection, which simplifies the analysis and visualization of complex patterns of viral mutations during the course of an infection. It also allows for a statistical analysis of the relationship between viral mutations and the immune response targeting specific HCV variants. This type of software package is likely to become increasingly common in the near future, as a result of the increasingly large amount of data being rapidly generated and the overwhelming need for computational tools for analysis of complex multidisciplinary data in a time efficient manner.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This project was supported by National Health and Medical Research Council of Australia (NHMRC) Program nos. 1060199, 510448, and 1042090 (to Fabio Luciani, Rowena Bull, and Andrew Lloyd), and by grants from Australian Centre for HIV and Hepatitis Virology (ACH2). Rowena Bull and Andrew Lloyd are supported by NHMRC Fellowships no. 630733 and no. 1043067. The HITS-p investigators include Andrew Lloyd, Kate Dolan, Paul Haber, William Rawlinson, Carla Treloar, Greg Dore, Lisa Maher, and Fabio Luciani.

References

[1] M. Rolland, S. Tovanabutra, A. C. deCamp et al., "Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial," *Nature Medicine*, vol. 17, no. 3, pp. 366–371, 2011.

[2] K. McElroy, T. Thomas, and F. Luciani, "Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions.," *Microbial Informatics and Experimentation*, vol. 4, article 1, 2014.

[3] F. Luciani, R. A. Bull, and A. R. Lloyd, "Next generation deep sequencing and vaccine design: today and tomorrow," *Trends in Biotechnology*, vol. 30, no. 9, pp. 443–452, 2012.

[4] M. J. Clemente, B. Przychodzen, A. Jerez et al., "Deep sequencing of the T-cell receptor repertoire in CD8+ T-large granular lymphocyte leukemia identifies signature landscapes," *Blood*, vol. 122, no. 25, pp. 4077–4085, 2013.

[5] D. A. Bolotin, I. Z. Mamedov, O. V. Britanova et al., "Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms," *European Journal of Immunology*, vol. 42, no. 11, pp. 3073–3083, 2012.

[6] R. J. Bashford-Rogers, A. L. Palser, B. J. Huntly, R. Rance, G. S. Vassiliou et al., "Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations," *Genome Research*, vol. 23, no. 11, pp. 1874–1884, 2013.

[7] R. A. Bull, F. Luciani, K. McElroy et al., "Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection," *PLoS Pathogens*, vol. 7, no. 9, Article ID e1002243, 2011.

[8] F. Di Giallonardo, O. Zagordi, Y. Duport, C. Leemann, B. Joos et al., "Next-generation sequencing of HIV-1 RNA genomes: determination of error rates and minimizing artificial recombination," *PLoS ONE*, vol. 8, no. 9, Article ID e74249, 2013.

[9] A. Plauzolles, M. Lucas, and S. Gaudieri, "Hepatitis C virus adaptation to T-cell immune pressure," *The Scientific World Journal*, vol. 2013, Article ID 673240, 7 pages, 2013.

[10] J. Grebely, K. Page, R. Sacks-Davis, M. S. van der Loeff, T. M. Rice et al., "The effects of female sex, viral genotype, and IL28B genotype on spontaneous clearance of acute hepatitis C virus infection," *Hepatology*, vol. 59, no. 1, pp. 109–120, 2014.

[11] A. Ascione, T. Tartaglione, and G. G. Di Costanzo, "Natural history of chronic hepatitis C virus infection," *Digestive and Liver Disease*, vol. 39, supplement 1, pp. S4–S7, 2007.

[12] M. S. Choi, D. Y. Kim, D. H. Lee et al., "Clinical significance of pre-S mutations in patients with genotype C hepatitis B virus infection," *Journal of Viral Hepatitis*, vol. 14, no. 3, pp. 161–168, 2007.

[13] R. Sanjuan, A. Moya, and S. F. Elena, "The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 22, pp. 8396–8401, 2004.

[14] J. W. Drake and J. J. Holland, "Mutation rates among RNA viruses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 24, pp. 13910–13913, 1999.

[15] S. F. Elena and R. Sanjuan, "Virus evolution: insights from an experimental approach," *Annual Review of Ecology, Evolution, and Systematics*, vol. 38, pp. 27–52, 2007.

[16] S. C. Manrubia, C. Escarmis, E. Domingo, and E. Lazaro, "High mutation rates, bottlenecks, and robustness of RNA viral quasispecies," *Gene*, vol. 347, no. 2, pp. 273–282, 2005.

[17] S. C. Manrubia and E. Lázaro, "Viral evolution," *Physics of Life Reviews*, vol. 3, no. 2, pp. 65–92, 2006.

[18] D. G. Bowen and C. M. Walker, "Adaptive immune responses in acute and chronic hepatitis C virus infection," *Nature*, vol. 436, no. 7053, pp. 946–952, 2005.

[19] P. Klenerman and R. Thimme, "T cell responses in hepatitis C: the good, the bad and the unconventional," *Gut*, vol. 61, no. 8, pp. 1226–1234, 2012.

[20] J. Neefjes, M. L. Jongsma, P. Paul, and O. Bakke, "Towards a systems understanding of MHC class I and MHC class II antigen presentation," *Nature Reviews Immunology*, vol. 11, no. 12, pp. 823–836, 2011.

[21] C. Oniangue-Ndza, T. Kuntzen, M. Kemper et al., "Compensatory mutations restore the replication defects caused by cytotoxic T lymphocyte escape mutations in hepatitis C virus polymerase," *Journal of Virology*, vol. 85, no. 22, pp. 11883–11890, 2011.

[22] H. Li, M. B. Stoddard, S. Wang et al., "Elucidation of hepatitis C virus transmission and early diversification by single genome sequencing," *PLoS Pathogens*, vol. 8, no. 8, Article ID e1002880, 2012.

[23] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589–595, 2010.

[24] H. Li, B. Handsaker, A. Wysoker et al., "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[25] O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel, "ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data," *BMC Bioinformatics*, vol. 12, article 119, 2011.

[26] S. Rhee, T. F. Liu, S. P. Holmes, and R. W. Shafer, "HIV-1 subtype B protease and reverse transcriptase amino acid covariation," *PLoS Computational Biology*, vol. 3, no. 5, article e87, 2007.

[27] M. Krzywinski, J. E. Schein, I. Birol et al., "Circos: an information aesthetic for comparative genomics," *Genome Research*, vol. 19, no. 9, pp. 1639–1645, 2009.

[28] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, and A. Bairoch, "ExPASy: the proteomics server for in-depth protein knowledge and analysis," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3784–3788, 2003.

[29] I. Kass and A. Horovitz, "Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations," *Proteins*, vol. 48, no. 4, pp. 611–617, 2002.

[30] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[31] R. Aurora, M. J. Donlin, N. A. Cannon, and J. E. Tavis, "Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans," *The Journal of Clinical Investigation*, vol. 119, no. 1, pp. 225–236, 2009.

[32] A. Wilm, P. P. Aw, D. Bertrand et al., "LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets," *Nucleic Acids Research*, vol. 40, no. 22, pp. 11189–11201, 2012.

[33] S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0," *Systematic Biology*, vol. 59, no. 3, pp. 307–321, 2010.

[34] N. Beerenwinkel, H. F. Gunthard, V. Roth, and K. J. Metzner, "Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data," *Frontiers in Microbiology*, vol. 3, article 329, 2012.

[35] K. McElroy, O. Zagordi, R. Bull, F. Luciani, and N. Beerenwinkel, "Accurate single nucleotide variant detection in viral populations by combining probabilistic clustering with a statistical test of strand bias," *BMC Genomics*, vol. 14, article 501, 2013.

[36] M. Schirmer, W. T. Sloan, and C. Quince, "Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes," *Briefings in Bioinformatics*, vol. 15, no. 3, pp. 431–432, 2012.

[37] A. J. McMichael, P. Borrow, G. D. Tomaras, N. Goonetilleke, and B. F. Haynes, "The immune response during acute HIV-1 infection: clues for vaccine development," *Nature Reviews Immunology*, vol. 10, no. 1, pp. 11–23, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 731325, 13 pages
http://dx.doi.org/10.1155/2014/731325

Research Article

Finding Semirigid Domains in Biomolecules by Clustering Pair-Distance Variations

Michael Kenn,[1] Reiner Ribarics,[1] Nevena Ilieva,[2] and Wolfgang Schreiner[1]

[1] *Section of Biosimulation and Bioinformatics, Center for Medical Statistics, Informatics, and Intelligent Systems (CeMSIIS), Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria*
[2] *Institute for Nuclear Research and Nuclear Energy (INRNE), Bulgarian Academy of Sciences, 72, Tzarigradsko Chaussee, 1784 Sofia, Bulgaria*

Correspondence should be addressed to Wolfgang Schreiner; wolfgang.schreiner@meduniwien.ac.at

Received 31 January 2014; Accepted 10 March 2014; Published 15 May 2014

Academic Editor: Francesco Pappalardo

Dynamic variations in the distances between pairs of atoms are used for clustering subdomains of biomolecules. We draw on a well-known target function for clustering and first show mathematically that the assignment of atoms to clusters has to be crisp, not fuzzy, as hitherto assumed. This reduces the computational load of clustering drastically, and we demonstrate results for several biomolecules relevant in immunoinformatics. Results are evaluated regarding the number of clusters, cluster size, cluster stability, and the evolution of clusters over time. Crisp clustering lends itself as an efficient tool to locate semirigid domains in the simulation of biomolecules. Such domains seem crucial for an optimum performance of subsequent statistical analyses, aiming at detecting minute motional patterns related to antigen recognition and signal transduction.

1. Introduction

Molecular dynamics (MD) can be used to investigate functional elements in biomolecules [1–5]. In addition to static structures (such as crystal structures stored in the protein data bank (PDB) [6]) molecular dynamics yields information on dynamic properties [7, 8], lending themselves for evaluation of, for example, signal transduction. However, key patterns of motion related to such a functional element may be hidden among a large amount of "other" movements, reflecting no more than ordinary thermal motility of the biomolecule. Molecular dynamics itself can be carried out along relatively standardized protocols [9, 10]. However, recognizing specific patterns of motion, which are deemed crucial for a functional element, remains a tricky task, requiring sophisticated statistical methods [11], such as principal component analysis [12, 13] or normal mode analysis [14].

For all the mentioned approaches, an initial and essential step is the "fitting" of the molecular structure of each time step of an MD trajectory (henceforward called frame) to a reference structure, $\mathbf{x}_{\text{ref}}$ [15]. A given frame $\mathbf{x}_i$ is first translated to let its centre of mass coincide with that of the reference frame. Then $\mathbf{x}_i$ is rotated (around its centre of mass) to minimize square deviations between corresponding atoms of $\mathbf{x}_i$ and $\mathbf{x}_{\text{ref}}$:

$$\text{RMSD}\left(\mathbf{x}_i, \mathbf{x}_{\text{ref}}\right) = \left[\frac{1}{\sum_i w_i} \sum_{i=1}^{N} w_i \left\|\mathbf{x}_i - \mathbf{x}_{\text{ref}}\right\|^2\right]^{1/2} \longrightarrow \min. \tag{1}$$

In many approaches, RMSD has been used not only for fitting but also for directly (and successfully) quantifying molecular deformations [15], including structural changes, drifts, and trends [16]. In many cases, however, even sophisticated statistical methods fail, when applied to MD-frames after fitting. The suspicion is that the process of fitting itself might cause this failure. How does this come about?

By default, the GROMACS [17] fitting procedure uses atomic masses as weights for superposition of a structure's atomic coordinates to a reference structure. Accordingly, the fitting of $\mathbf{x}_i$ "as a whole" is being optimized. In some cases, fitting the whole molecule may be inadequate and even conceal what one is searching for. For example, consider a molecule with one or more flexible loops. While the body of such a molecule behaves like a slightly deformable, rigid body, a loop may be conformationally flexible exhibiting

largely uncorrelated movements with respect to the rest of the molecule. In the fitting criterion, however, atoms within the loop and those in the body may have equal weights. Since all deviations enter quadratic into (1), large movements of an even small number of loop-atoms may generate dominant contributions to the RMSD. In such a case, in order to minimize total RMSD, $\mathbf{x}_i$ is rotated predominantly to accommodate for the few atoms within the loop. As a result, the large remaining body of the molecule has to "follow its own loop," as if the tail chases the dog [maybe this was not the primary intention of fitting]. Needless to say, due to such movements caused by fitting, that minute motile elements may become totally submerged, without any chance of being retrieved from the trajectory, not even by sophisticated statistics.

The described situation is typical and demands more elaborate fitting methods. Choosing unequal weights suggests itself as a nearby and convenient solution. The more rigid parts of the molecule should receive more weight, the flexible ones less. However, how should one know, prior to fitting, which parts are semirigid and which are flexible?

One possibility would be a two-pass procedure, in the first pass fitting to the whole molecule with uniform weights ($w_i = 1$) and evaluating the RMSF_i:

$$\mathrm{RMSF}_i = \sqrt{\left\langle \left(\mathbf{x}_i - \mathbf{x}_{i,\mathrm{ref}}\right)^2 \right\rangle}, \tag{2}$$

where $\langle\rangle$ denotes the average over a trajectory and $\mathbf{x}_{i,\mathrm{ref}}$ denotes a reference position of atom i, not changing over time. Note that RMSF_i will highly depend on the choice of the reference position, which is usually the mean coordinate of atom i over the whole trajectory. Then, in a second pass of fitting, weights are chosen inversely proportional to the RMSF_i, as reported by [18]. Highly motile atoms receive less weight and lose their role in shaking the remaining main parts of the molecule. However, this method suffers from the fact that RMSF_i depends on the selection of $\mathbf{x}_{i,\mathrm{ref}}$ in the first pass of fitting; that is, the correction procedure depends on the error it is supposed to correct.

Another possibility is the identification of semirigid domains (clusters) within the molecule, as reported by [18]. In particular, the definition of clusters may be based on distances $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ between pairs of atoms rather than coordinates computed in the trajectory. The standard deviation of distance variation (STDDV) between an atomic pair (i, j) is given by

$$S_{ij} = \sqrt{\frac{N}{N-1}\left\langle \left(d_{ij} - \left\langle d_{ij}\right\rangle\right)^2 \right\rangle}, \tag{3}$$

where N is the number of atoms, $\langle\rangle$ denotes the average over a trajectory, and S_{ij} is measured in nm. Evidently, pair-distances are unaffected by any kind of arbitrariness due to fitting.

Given a number of clusters (N_{clust}), let c_{im} denote the partial class membership of atom i in cluster m. For normalization we require

$$\sum_{k=1}^{N_{\mathrm{clust}}} c_{im} = 1. \tag{4}$$

Table 1: Molecular complexes simulated.

Molecular system	Simulation length
Penta-L-alanine (A_5)	1000 ns
LC13 TCR/ABCD3/HLA-B*44:02 (B4402)	250 ns
LC13 TCR/ABCD3/HLA-B*44:03 (B4403)	250 ns

The following criterion has been proposed to identify an optimum decomposition into a given number (N_{clust}) of clusters [18]. Minimize the target function:

$$q(\mathbf{c}) = \sum_{m=1}^{N_{\mathrm{clust}}}\sum_{i=1}^{N}\sum_{j=1}^{N} c_{im}c_{jm}S_{ij} = \mathrm{tr}\left(\mathbf{c}^T\mathbf{S}\mathbf{c}\right) \longrightarrow \min \tag{5}$$

under the constraints of (4). Once identified, any such cluster may be used as "primary fitting domain," by assigning large weights to the atoms therein. With little motion within such a cluster, the motility of the remaining atoms of the molecule will appear relative to that cluster. This generally increases the chance of tracing relevant patterns of motion outside the cluster, for many statistical methods being applied.

The important difference from the known structure-analyzing tools of GROMACS, whether based on RMS deviation after fitting or RMS deviation of atom-pair distances, is that there is no need of a reference structure here. Also the clustering algorithms themselves, though with different criteria, assign conformations to a cluster for the molecule as a whole, while in our work, groups of atoms are assigned to the same cluster if their mutual distances vary little over time (a spatial clustering within the molecule).

MD trajectories for protein complexes were analyzed by clustering of averaged standard deviations of distance variation (STDDV); see below. Obtaining the most rigid cluster of atoms can be seen as the first step to facilitate the search for protein motions.

2. Methods

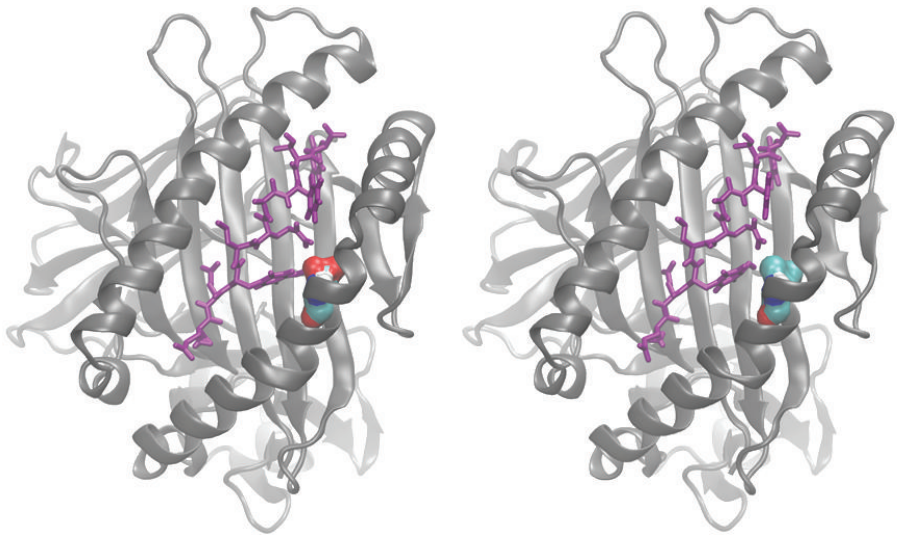
2.1. Construction of Complexes for Molecular Dynamics Simulation. We applied the clustering algorithm to a series of molecular systems as follows, see Table 1.

LC13 T cell receptor (TCR) in complex with major histocompatibility complex (MHC) HLA-B*44:05 and the ABCD3 peptide (EEYLQAFTY) has been successfully crystallized by Macdonald et al. [20] and its structure is accessible on http://www.pdb.org/ assigned the PDB ID *3KPS*. However, there are no structure files of LC13 TCR in complex with HLA-B*44:02 and HLA-B*44:03. Therefore, we applied homology modelling to create these structures.

In-silico mutagenesis was carried out using Swiss PDB Viewer [21]. For generation of LC13/ABCD3/HLA-B*44:03, we used PDB structure *3KPS* as a template and introduced mutations Y116D (numbers according to PDB numbering) and D156L to the MHC thus changing the HLA type from B*44:05 to B*44:03. For generation of LC13/ABCD3/HLA-B*44:02 we used PDB structure *3KPS* as a template and introduced mutation Y116D to the MHC thus changing the HLA type from B*44:05 to B*44:02; see Figures 1 and 2.

HLA numbering	140	160	180
	|	|	|
PDB numbering	120	140	160
	|	|	|
B44:05/130–200	DGRLLRGYDQYAYDGKDYIALNEDLSSWTAADTAAQITQRKWEAARVAEQDRAYLEGLCVESLRRYLENGK		
B44:03/130–200	DGRLLRGYDQDAYDGKDYIALNEDLSSWTAADTAAQITQRKWEAARVAEQLRAYLEGLCVESLRRYLENGK		
B44:02/130–200	DGRLLRGYDQDAYDGKDYIALNEDLSSWTAADTAAQITQRKWEAARVAEQDRAYLEGLCVESLRRYLENGK		

Figure 1: Alignment of amino acid sequences of HLA-B*44:02, HLA-B*44:03, and HLA-B*44:05 (downloaded from IMGT/HLA database [19]). HLA-B*44:05 was used as a template for homology modeling, because a three-dimensional structure of this MHC in complex with ABCD3 peptide and LC13 TCR was available. Sequence alignment was done with CLC bio's CLC sequence viewer. Note that sequence numbering from PDB (PDB numbering) and IMGT/HLA database (HLA numbering) differ.



(a) (b)

Figure 2: Visualisation of the D156L mutation in the MHC molecule. MHC molecules (gray) HLA-B*44:02 (left), and HLA-B*44:03 (right) together with ABCD3 peptide (violet) are shown. The amino acids comprising the D/L polymorphisms at position 156 are shown in surface representation (red···oxygen, blue···nitrogen, turquoise···carbon, white···hydrogen). Parts of the ABCD3 peptide closely interact with residue 156(D/L).

2.2. Molecular Dynamics Simulation Protocol. The workflow of the molecular dynamics simulation of the penta-L-alanine system is closely related to that in the work of Bernhard and Noé [18]. MD simulation of penta-L-alanine was performed using GROMACS 4.0.7 [17] according to the following protocol.

First, we immersed penta-L-alanine in an explicit SPC [22] artificial water bath (cubic box) allowing for a minimum distance of 1 nm between peptide and box boundaries. Second, we minimized the solvated system using a steepest descent method. Next, we warmed up the system to 293 K during a 100 ps position restraint MD simulation. Finally, we carried out the MD production run with LINCS constraint algorithm acting on bonds with hydrogen atoms using an integration step of 2 fs and the GROMOS96 53a6 force field [23]. Coordinates were written to the trajectory every 2 ps. Coulomb interactions were computed using Particle Mesh Ewald (PME) with a maximum grid spacing of 0.12 nm and interpolation order 4. Both, Van der Waals and Coulomb interactions were computed with a cut-off at 1.4 nm. Berendsen temperature coupling to 293 K and Berendsen isotropic pressure coupling to 1 bar were used. All further parameters were set in accordance with Omasits et al. [24].

MD simulation of TCR/pMHC systems was performed using GROMACS 4.0.7 [17] according to the following protocol. First, we immersed the TCR/pMHC complex in SPC [22] artificial water bath (cubic box) allowing for a minimum distance of 2 nm between complex and box boundaries. Second, we added sodium and chloride ions to a concentration of 0.15 mol/L, and at the same time neutralizing the net charge of the system. Third, we minimized the energy of the solvated system using a steepest descent method. Next, we warmed up the system to 310 K during a 100 ps position restraints MD simulation. Finally, we carried out MD production runs with LINCS constraint algorithm acting on all bonds and using the GROMOS96 53a6 force field [23]. Hydrogen motions were removed allowing for an integration step of 5 fs. Coordinates were written to the trajectory every 50 ps. Coulomb interactions were computed using Particle Mesh Ewald (PME) with a maximum grid spacing of 0.12 nm and interpolation order 4. Both, Van der Waals and Coulomb interactions were computed with a cut-off at 1.4 nm. Velocity rescale temperature coupling to 310 K and Berendsen isotropic pressure coupling to 1 bar were used. All further parameters were set in accordance with Omasits et al. [24].

2.3. Optimization of Cluster Membership. Each atom i in a molecular dynamics simulation may be uniquely assigned to one of the N_{clust} clusters considered [7, 25, 26], represented by a "crisp" vector of cluster-membership; for example, $c_{im} = [0, 0, 0, 1, 0, 0]$ if atom i belongs to cluster 4 out of $N_{\text{clust}} = 6$ clusters. Alternatively, each atom i may be considered to belong to several clusters simultaneously, represented by fuzzy, noninteger memberships, with normalization condition see (4). Fuzzy memberships are the more general case, it seems that they might yield lower minima of the target function than crisp memberships and should therefore be preferred. Interestingly, Bernhard and Noé [18] report that fuzzy memberships, upon optimization with a gradient method, tend to end up as crisp, that is, either 0 or 1. We have scrutinized this issue and will demonstrate how this comes about. Even more, as one of the main results of this work, we will prove that the solution has to be crisp. The proof is given via mathematical arguments; see results section. This finding allows us to restrict the search space to crisp memberships, without diminishing the generality of the optimization problem posed.

2.3.1. Optimization of Crisp Cluster Memberships by a Two-Stage Monte Carlo Method. Initially, the number of clusters, N_{clust}, is chosen and the target function (5) has to be minimized. We will show that, under certain assumptions (see Section 3.1), it is sufficient to search in

$$\Omega^* = \left\{ c_{im} \mid 1 \le i \le N, 1 \le m \le N_{\text{clust}}, c_{im} \in \{0, 1\}, \sum_{m=1}^{N_{\text{clust}}} c_{im} = 1 \right\}. \tag{6}$$

In a less formal formulation, the objective is to assign each of the N atoms to one particular cluster (crisp memberships).

It may become quite tricky to attack such a problem with an analytic gradient approach since the boundary conditions are usually difficult to handle and the search domain consists of isolated points. We have therefore chosen a two-step search, in which a random process is succeeded by an exhaustive search.

Every constellation (i.e., Monte Carlo trial for improvement) which cannot be improved by a single move of one of the N atoms from one cluster to another is considered a result (minimum constellation). The result with the lowest $q(\mathbf{c})$ is the ground state, but all other minimum constellations should also be included in the further analysis.

2.3.2. Search Algorithm

(i) Start-Up. Initially, each of the N components is randomly assigned to one (of the N_{clust}) cluster.

(ii) Random Search. In the first step, each of the N components (atoms) is moved from its current cluster to another randomly chosen cluster with probability P. If this mutation yields a reduction in $q(\mathbf{c})$ the new constellation is preserved; otherwise, it is rejected. This process is repeated K times. A very rudimentary benchmarking analysis has shown that $P = 1/N$ and $K = N \cdot N_{\text{clust}}$ are reasonable values to use.

(iii) Exhaustive Search. In the second step, the algorithm tries to improve $q(\mathbf{c})$ by single step moves for each component separately. If there is no possible move to improve $q(\mathbf{c})$, the constellation is necessarily a local minimum in our sense.

(iv) Ground State. Usually there is a large number of minimum constellations in the above sense. For a matrix with significant structure, the ground state will be reached after only a few trials. For ill-conditioned matrices (those without structure), a minimum constellation very close to the ground state will also be found after a few trials, although the absolute ground state might be difficult to find.

3. Results

3.1. Crisp Cluster Membership as a Necessary Consequence. We formulate and prove a lemma that crisp memberships are a necessary consequence of the topology of the multidimensional space of pair-distance standard deviations.

Due to its definition, $\mathbf{S}$ is a symmetric, nonsingular $N \times N$ adjacency matrix whose entries are the standard deviations explained earlier (3); thus, $S_{ij} > 0$ for $i \neq j$ and $S_{ii} = 0$. Let

$$\Omega = \left\{ c_{im} \mid 1 \le i \le N, 1 \le m \le N_{\text{clust}}, c_{im} \ge 0, \sum_{m=1}^{N_{\text{clust}}} c_{im} = 1 \right\}. \tag{7}$$

The objective is to find $\widehat{\mathbf{c}} = \operatorname{argmin}\, q(\mathbf{c})$ for $\mathbf{c} \in \Omega$.

3.1.1. Lemma. If $\mathbf{S}$ is a symmetric, nonsingular $N \times N$ matrix with nonnegative entries and $\widehat{\mathbf{c}} = \operatorname{argmin}\, q(\mathbf{c})$, for $\mathbf{c} \in \Omega$, then $\widehat{c}_{im} \in \{0, 1\}$.

To prove this lemma one uses Lagrange multipliers:

$$q\left(c_{11}, \ldots, c_{N \cdot N_{\text{clust}}}, \lambda_1, \ldots, \lambda_N\right) = \sum_{m=1}^{N_{\text{clust}}} \sum_{i=1}^{N} \sum_{j=1}^{N} c_{im}\, c_{jm}\, S_{ij} + \sum_{i=1}^{N} \lambda_i \left(\sum_{m=1}^{N_{\text{clust}}} c_{im} - 1 \right) \longrightarrow \min. \tag{8}$$

Since $q(\mathbf{c})$ is a polynomial of order 2, the derivatives with respect to c_{im} and λ_i yield a system of $N \times (N_{\text{clust}} + 1)$ linear equations of the form:

$$\begin{pmatrix} S & 0 & \cdots & 0 & I_N \\ 0 & S & \cdots & 0 & I_N \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & S & I_N \\ I_N & I_N & \cdots & I_N & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \cdots \\ \mathbf{c}_{N_{clust}} \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ 1 \end{pmatrix} \tag{9}$$

with I_N being $N \times N$ identity matrix and $\mathbf{c}_m = (c_{1m}, \ldots, c_{Nm})$. The determinant of the matrix is $-N_{\text{clust}} \det(\mathbf{S})^{N_{\text{clust}}-1}$ and

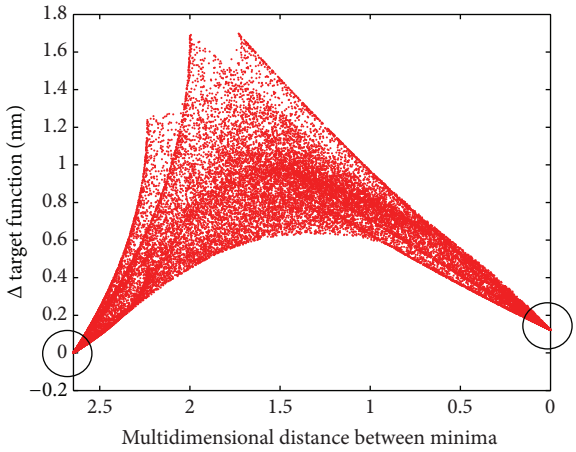


FIGURE 3: Topology of target function if a group of 7 atoms is allowed to switch between different clusters. At left and right edge of the graphics ground states are located (black circles), with corresponding values of the target function (reference minima, left somewhat lower than right). For the left state, all 7 switchable atoms are located according to cluster minimum 1, for the right state according to cluster minimum 2. In between, target function values are plotted (as red dots) for all permutations of cluster memberships for the 7 switchable atoms (3 degrees of freedom: 2 + 2 + 3 atoms). Due to the exceedingly high number of permutations, the area is densely filled with dots. Vertical axis: target function [nm] relative to minimum shown at left margin. Horizontal axis: multidimensional Euclidean distance between reference minima.

therefore, under the assumption $\det(\mathbf{S}) \neq 0$, there must be a unique solution, which is given by

$$c_{im} = \frac{1}{N_{\text{clust}}} \qquad \lambda_i = -\frac{2}{N_{\text{clust}}} \sum_{j=1}^{N} S_{ij}. \tag{10}$$

Unfortunately, this solution yields the maximum of $q(\mathbf{c})$ for $\mathbf{c} \in \Omega$. However, since Ω is convex and bounded, one knows that any argmin $q(\mathbf{c})$ must be on $\delta\Omega$. Therefore, there must be at least one $c_{i'm'} = 0$. Reducing the above system of linear equations by this constraint, one sees that the revolving system again has a unique solution:

$$c_{im'} = \frac{1}{N_{\text{clust}} - 1} \quad \text{for } m \neq m' \qquad c_{im} = \frac{1}{N_{\text{clust}}} \quad \text{for } i \neq i'$$

$$\lambda_i = -\frac{2}{N_{\text{clust}}} \sum_{j=1, j\neq i}^{N} S_{i,j} \tag{11}$$

which again gives a maximum of $q(\mathbf{c})$. The derivation of the value of λ_i is rather involved and not shown.

With the same argument as before, one can proceed by setting all c_{im} equal to zero with the exception of one c_{im} for each m. This iterative procedure shows that clustering with respect to $q(\mathbf{c})$ leads to a unique assignment of each of the N components to one particular cluster.

3.2. Heuristic Evaluation of Solution Space around the Minima. Our theoretical result, that memberships are crisp, can be illustrated very intuitively; see Figure 3. Left and right end of x-axis correspond to constellations with minimum target function, located at the boundary, and no other minimum is found in between.

3.3. Clusters of Atomic Motions in MD Trajectory. For above mentioned TCR/pMHC complexes B4402 and B4403, STDDV matrices were computed from the MD trajectories; see Figure 4. Only the second parts of the trajectories (corresponding to approx. 125 ns simulation time) were considered to exclude relaxation effects; see also Section 3.6.1.

STDDV matrices were clustered as described above for $N_{\text{clust}} = 2$ to 6. After computation, clusters are renumbered according to size (=number of atoms), the largest one always being labelled as cluster 1. For $N_{\text{clust}} = 5, 6$ cluster memberships for B4402 and B4403 were remapped onto the protein structure and displayed in VMD [27]; see Figure 5.

Note that NO information whatsoever about secondary structural elements, such as α-helices and β-sheets, has entered the clustering procedure. Still, clusters more or less seem to retrieve some of these structural elements; see Figure 5. This could be related to extensive hydrogen bonding in α-helices and β-sheets stabilizing these secondary protein structure elements. How could bond constraints in MD simulations influence the resulting clusters? In our calculations, we just considered the protein backbone, because amino acids side chains show larger spatial fluctuations. The backbone C_α atoms are separated by a planar and rigid amid bond, so neighboring C_α atoms will experience less variation in distance.

3.4. More Clusters Improve Target Function. The number of clusters has to be preselected in our approach. If we had just one cluster, the total distance variability contained in matrix $\mathbf{S}$ would be part of that cluster. Increasing the number of clusters generally reduces the fraction of variability contained within clusters, expressed as percentage of total in Figure 6.

3.5. Larger Clusters Turn Out to Be More Rigid. Clusters were constructed to achieve maximum internal "rigidity," that is, a minimum sum of pair-distance standard deviations. One might expect that large clusters, since they accommodate many atoms within larger spatial domains, should turn out to be less rigid then smaller clusters. However, the opposite is true: larger clusters turn out to be more rigid; see the declining trend of normalized STDDV with increasing cluster size in Figure 7. This demonstrates again that structures in motility are captured via clustering. If there is no structure within matrix $\mathbf{S}$, normalized cluster sizes would result nearly equal, that is, centered around 1.

Clearly, clusters do not result identical for different subsections of a MD trajectory. In Figure 7, data are shown for two trajectories and 50 subsections each, each clustered for $N_{\text{clust}} = 5$ and 6. For details, see legend of Figure 7. 100 Monte Carlo attempts were performed for each clustering, out of which the optimum (smallest target function) was adopted. These results confirm the general trend that larger clusters are more rigid.

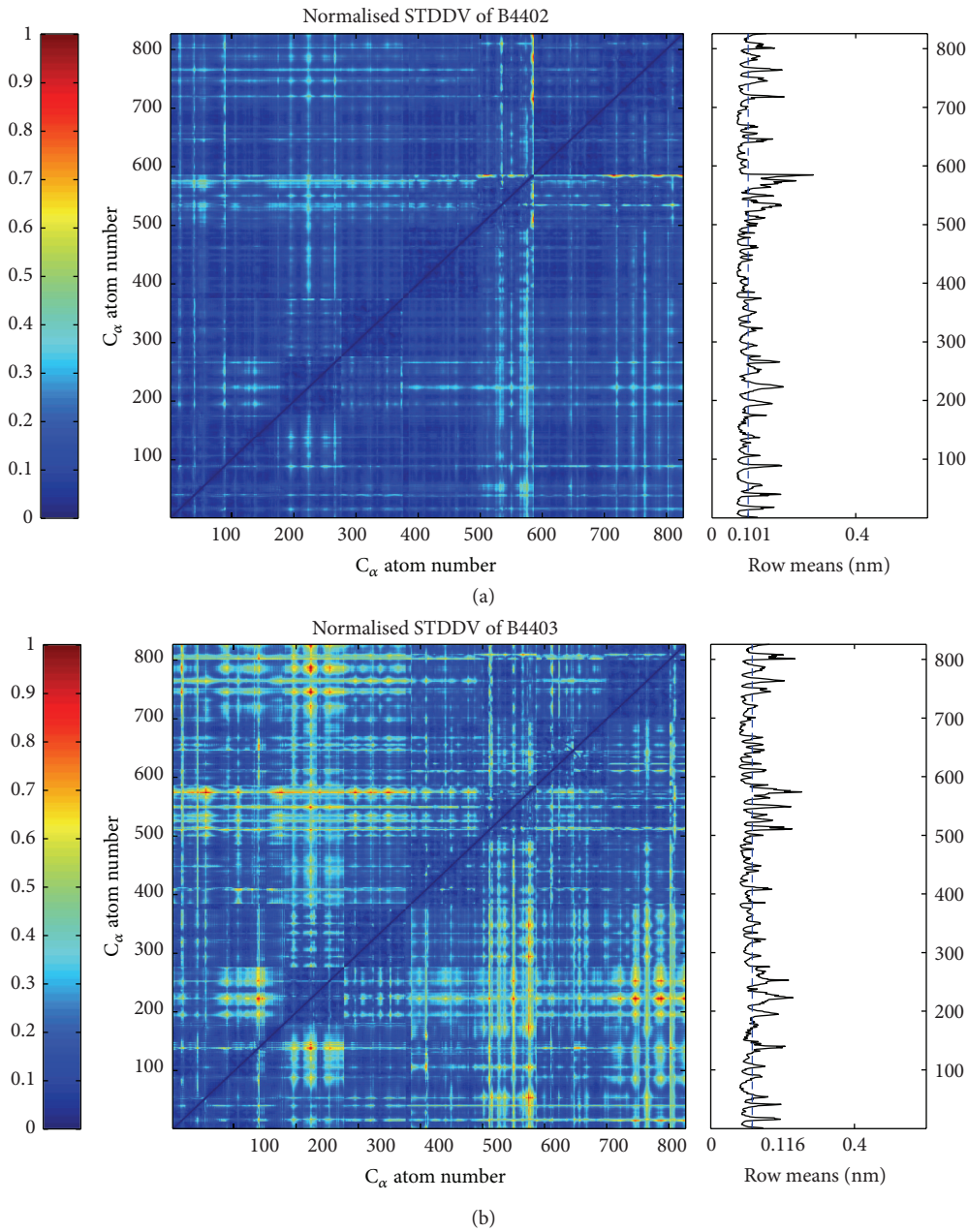


Figure 4: Standard deviation of pair distances (STDDV) in the second half of the trajectories of TCR/pMHC molecules B4402 (allogeneic) and B4403 (nonreactive). Values of STDDV [nm] have been normalized and are color coded (see bar on the left). Averaging over a row yields the mean distance variation against all other atoms (see subgraph on right). The dashed blue line shows the mean value of the row means, clearly indicating that the second half of the 250 ns trajectory of B4403 is more dynamic than B4402. Note that only C_α atoms are considered.

An overview of dispersions within and between clusters is given in Figure 8, the numerical results for 6 clusters being given in Table 2.

3.6. Stability of Clusters. Clusters have been evaluated regarding stability, in order to check whether they lend themselves as reliable semirigid domains for fitting MD-configurations. Of note that (at least) two sources of variability of cluster memberships need to be scrutinized as follows:

(i) variability due to the stochastic nature of the Monte Carlo clustering method and

(ii) variability due to different parts of an MD trajectory being clustered.

We will demonstrate that variability due to our Monte Carlo clustering method is negligible. As opposed to that, the "adequate" choice and preparation of the MD trajectory has tremendous impact and remains an issue of a never ending debate [28–32].

3.6.1. Variability between Different Parts of a Single MD Trajectory. Adequate sampling of phase space is essential regarding MD-simulations [33]. Much work has been done to

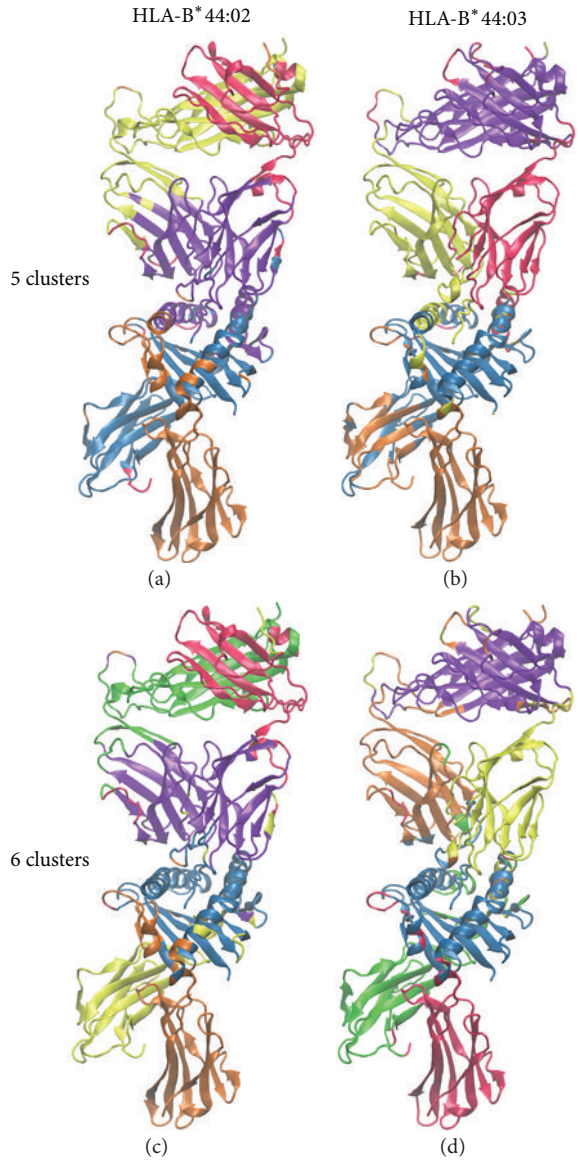


Figure 5: This 3D representation shows the LC13 TCR in complex with ABCD3 peptide and either HLA-B*44:02 (panels (a), (c)) or HLA-B*44:03 (panels (b), (d)). Number of clusters has been preset to five (upper panels) and six (lower panels). Clusters are rainbow-colored according to decreasing size (number of atoms in the cluster): violet (largest cluster), blue, green (relevant only for 6 clusters), yellow, orange, and red (the smallest cluster). The optimal clustering solution suggests that in these cases the most rigid clusters are the largest or the second largest ones (see also Figure 7). For panel (a) the most rigid cluster is blue and for panel (c) it is violet. For panel (b) the most rigid cluster is violet and for panel (d) it is blue. The most rigid cluster is therefore dependent on the prespecified number of clusters N_{clust}.

detect changes, drifts, and trends as markers for inadequate sampling [16, 29, 34]. Block averaging was proposed as one of the remedies [35].

In this work, the clustering presented above was based on matrices **S** computed from whole 250 ns MD trajectories. Clearly, the matrices $\mathbf{S}(t_1, t_2)$ for each subset (t_1, t_2) of a trajectory would be different, entailing different results for clustering. The question is which is the most reliable clustering for a given molecule?

To answer this question, we shall quantify the variability of clusters for subsets, relate it to the result for the whole trajectory, and derive a "stiff kernel," that is, those atoms which do not (or very rarely) change clusters between subsets of the trajectory.

Table 2: Dispersion of pair-distances within and between clusters.

	1	2	3	4	5	6
1	**0.058**	0.115	0.093	0.145	0.178	0.167
2	0.115	**0.068**	0.146	0.193	0.219	0.129
3	0.093	0.146	**0.079**	0.123	0.175	0.212
4	0.145	0.193	0.123	**0.075**	0.210	0.264
5	0.178	0.219	0.175	0.210	**0.090**	0.241
6	0.167	0.129	0.212	0.264	0.241	**0.124**

The full trajectory of protein complex B4402 was clustered into 6 clusters. Numbers in main diagonal give averaged STDDVs [nm] within clusters, corresponding to left markers (symbols: mean, minimum, maximum) in Figure 8. Off-diagonal values relate to STDDV between clusters, corresponding to right marker in Figure 8.

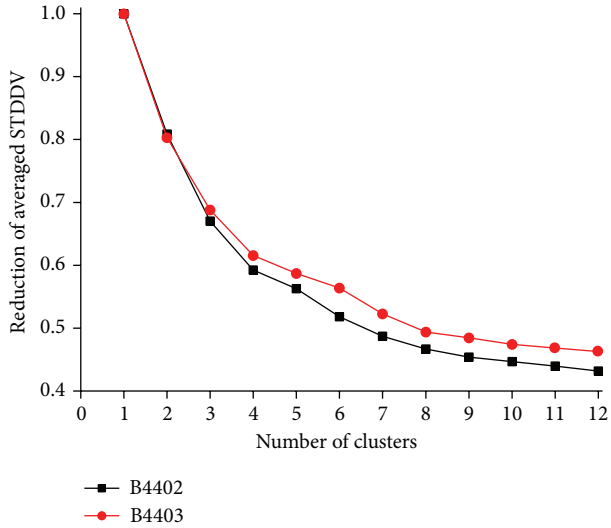


Figure 6: STDDV captured within clusters decreases with number of clusters. STDDV captured within clusters corresponds to the sum of areas of diagonal, coloured squares in Figure 13, right panel. Values of STDDV were normalized with respect to the total amount of STDDV without clustering ($N_{clust} = 1$) and have been multiplied by N_{clust} (i.e., number of squares) in order to be comparable. Note that the decline roots in the presence of mutual motile dependencies between pairs of atoms. If there were no depencencies, increasing the number of clusters would not significantly reduce STDDV, as demonstrated by the comparison with randomized dependencies shown in Figure 13.

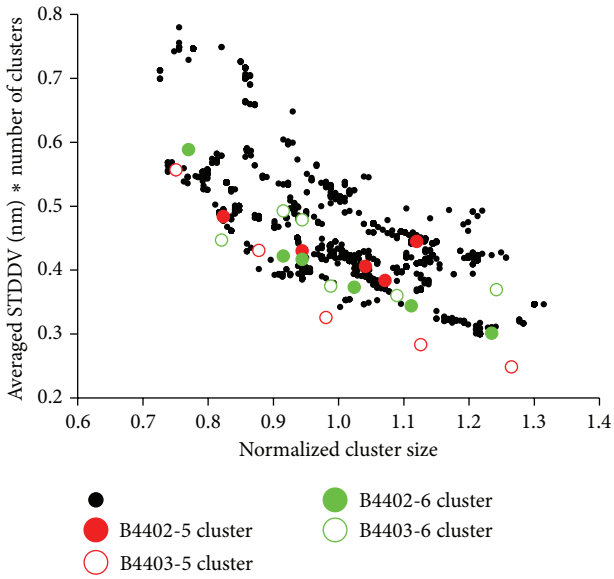


Figure 7: Rigidity versus size of clusters. Black data point refer to clusterings of two trajectories (B4402, B4403), each split into 50 subsections. For each subsection, clustering was performed twice, for $N_{clust} = 5$ and 6, yielding $2 \cdot 50 \cdot 2 = 200$ runs of clustering and $50 \cdot 2 \cdot (5 + 6) = 1100$ data points. Optimum solution for $N_{clust} = 5$ is shown in red, for $N_{clust} = 6$ in green (see Figure 5 for 3D visualization of clusters in the protein complexes). Vertical axis: STDDV within clusters was averaged and multiplied by N_{clust}, in order to be comparable. Horizontal axis: normalized cluster size = 1 means that the number of atoms in a cluster exactly matches the average N/N_{clust}.

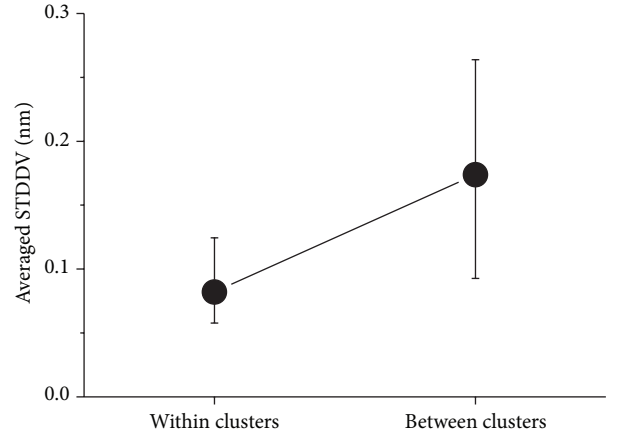


Figure 8: Standard deviation of pair distance variation (STDDV) within and between clusters. Symbols denote minimum, mean, and maximum.

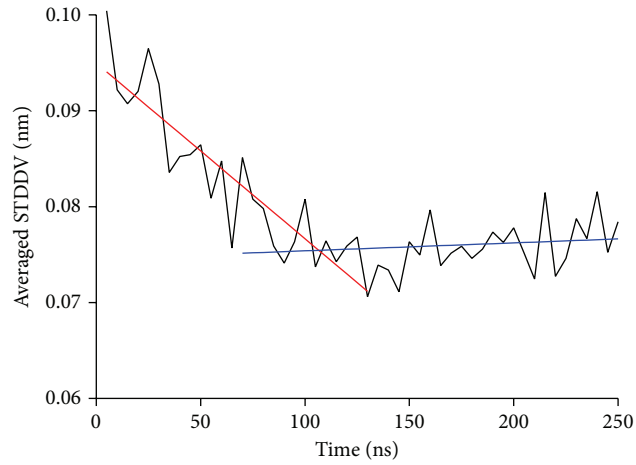


Figure 9: Time dependence of total variation, computed for subsets of a trajectory. Averaged STDDV (whole matrix $\mathbf{S}_{ij}$, example shown for B4402) computed for 50 subsets of trajectory, of 5 ns each. Straight lines have been fitted through time ranges 0–130 ns and 70–250 ns to illustrate a necessary discrimination between initial and ergodic phase of simulation.

First, we inspect the total dispersion contained within S_{ij}, evaluated separately for 50 subsets of 5 ns each (i.e., 100 frames out of 5000 frames in a whole trajectory); see Figure 9. In this trajectory we observe an irregular oscillating time behavior, starting with a declining tendency. The existence of such a substantial initial phase indicates that the influence of the starting configuration does not die out until after (roughly) half of the total simulation time has passed by. To account for this fact in a heuristic way, straight lines were fitted to the first and second half of the trajectory, allowing for some overlap. This accommodates with the finding of our previous work [29] that only the second half of the trajectory can be considered an unbiased sample from phase space and should be taken for further evaluations.

3.6.2. The Path along Most Stable Clusterings. For a preselected (number of clusters) N_{clust}, clustering was performed

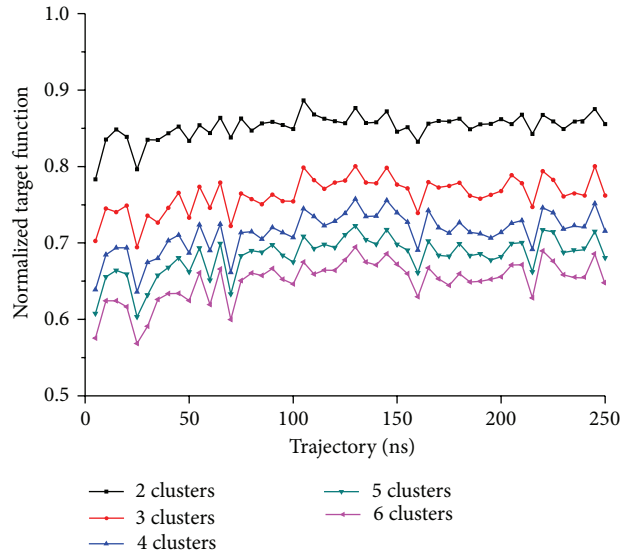


Figure 10: Time dependence of target function, evaluated for 50 subset-trajectories and different number of clusters. The total trajectory of 250 ns was split into 50 consecutive subset-trajectories and clustering performed for each of them, prescribing the number of clusters between 2 and 6 (see coloured legend). For each clustering, the resulting averaged STDDV is plotted against the vertical axis. Averaged STDDV have been multiplied by N_{clust} (as in Figure 7) to make them comparable between different numbers of clusters. Generally, more clusters entail smaller cluster size and thus reduce total motility captured in clusters. Variability of STDDV between subtrajectories illustrates the dynamical character of clustering and its dependence on the phase space sampling stage, however, with a pronounced convergence tendency of local values.

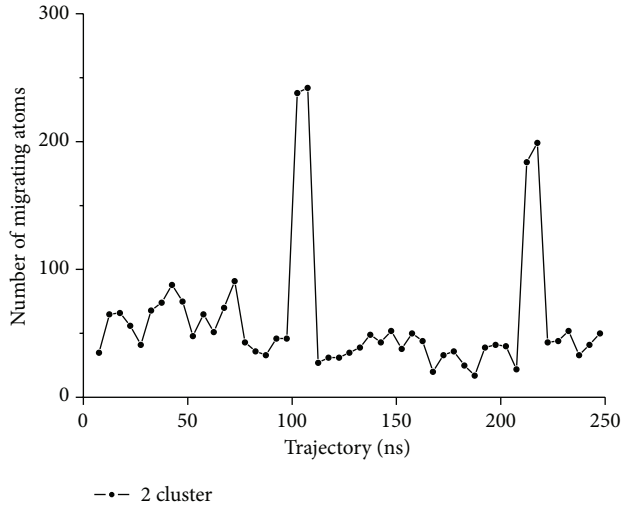


Figure 11: Migration of atoms between clusters. 50 subset-trajectories have been clustered (example shown for B4402, $N_{\text{clust}} = 2$). After relabelling (according to optimum permutation), the number of migrating atoms with respect to previous clustering is shown. Along this trajectory, two episodes of massive migration occur (around 100 and 220 ns, resp.). However, migrations turn out to be almost reversible; that is, most atoms finally end up in their former clusters (those of sub-trajectory 1), only about 50 (out of 826) do not.

for each of the 50 subset-trajectories (5 ns each), yielding an assignment for each atom to one of the clusters. Note that clusters were primarily labeled (cluster 1, 2, etc.) according to their size in that particular clustering (see Figure 10). Thus, the following situation may occur. Given a clustering of a subset-trajectory with first and second cluster about equal in size. Then, when clustering the following subset-trajectory, a few atoms from cluster 1 may end up in (i.e., "migrate" into) cluster 2, which may suffice to make cluster 2 now the largest cluster and therefore receiving the label "cluster 1" in the second clustering. This would yield a very peculiar result. The few "migrating atoms" would formally belong to the same cluster (in this example cluster 1), while the majority of atoms would switch between clusters 1 and 2. To avoid this misleading and undesired artifact, we refined the procedure as follows. In the first clustering, clusters were assigned labels (1, 2, etc.) according to decreasing size. After each subsequent clustering, we evaluated all permutations of cluster labels regarding the number of "migrating atoms" with respect to the first clustering. That permutation of labels, which yielded a minimum of migrating atoms, was finally adopted. As an example, the resulting number of migrating atoms is shown for B4402 and $N_{\text{clust}} = 2$ in Figure 11.

3.6.3. Quantifying the Stability of Cluster Assignments. In Figure 11 the number of migrating atoms was considered. Now, we evaluate which atoms migrate. For quantification, we resort to the Kullback-Leibler-distance [36, 37]:

$$\text{KLD}_i = \sum_{m=1}^{N_{\text{clust}}} p_{im} \cdot \log \frac{p_{im}}{1/N_{\text{clust}}}. \tag{12}$$

$1/N_{\text{clust}}$ represents the assumed background probability if the assignment of atom i to any of the clusters were equally probable. p_{im} is the actual probability for atom i to belong to cluster m, estimated from an average over cluster memberships c_{im} obtained from clustering subsets of the trajectory:

$$p_{im} = \langle c_{im} \rangle . \tag{13}$$

Large values of KLD_i indicate that atom i stays predominantly in the same cluster throughout the trajectory. On the contrary, values of KLD_i close to zero indicate a random distribution of an atom between all clusters. Figure 12 shows KLD_i for B4402 and $N_{\text{clust}} = 5$.

4. Discussion

4.1. Clustering Reflects Structure within STDDV-Matrix. Target function (5) only counts distance variability within the clusters, not between atoms belonging to different clusters. Thus, if we reorder atoms according to their cluster membership, clusters appear as squares along the main diagonal of the matrix **S**; see Figure 15. If we hypothetically assume that elements S_{ij} are more or less homogenously distributed across the matrix, the "area" of each cluster in the matrix will roughly correspond to the variability within that cluster. Clearly, these squares have to be of equal size to make their

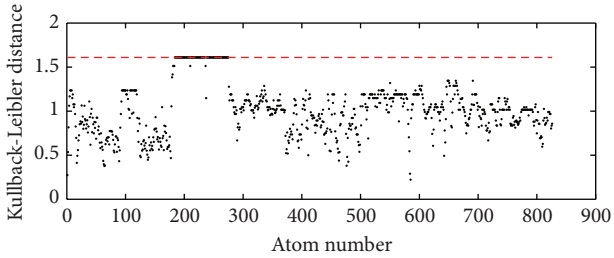


FIGURE 12: Kullback-Leibler distance as a measure for stability of cluster assignments among 50 subtrajectories. For each of the 50 subtrajectories, a separate clustering was performed. Comparing these clusterings, atoms are seen to move between the clusters several times. We consider the Kullback-Leibler distance (KLD) to estimate how far the distribution of the individual atoms among clusters deviates from a uniform distribution. Maximum possible values KLD = $\log_{10}(50)$, shown as red dashed line, correspond to atoms which never changed clusters. This can be observed, for example, for atoms with indices between 200 and 300. The lower the KLD, the more often the atoms move from one cluster to another. Data are shown for B4402, $N_{\text{clust}} = 5$.

joint area minimum (and thus minimize the total distance variation within clusters).

We have verified this prognosis by randomly rearranging elements of matrix **S** and then performing the clustering procedure. Cluster sizes resulted almost equal (for illustration see left panel of Figure 13) in each of 20 trials of random rearrangement and clustering. This finding was verified for $2 \leq N_{\text{clust}} \leq 6$ (data not shown).

As opposed, clusters of very different sizes resulted for matrices **S** derived from MD simulation (illustrated in right panel of Figure 13). This clearly indicates that clusters of unequal size do not result by chance but reflect distinct dependencies within **S**. The sensitivity of clustering to existing structures within **S** is also reflected in target function $q(\mathbf{S})$, as can be seen by comparing matrices **S** from MD and their randomly rearranged counterparts; see Figure 14. In both cases averaged STDDV decreases (improves) with increasing number of clusters. However, in presence of real dependencies between atomic mobility, clustering achieves much more improvement.

4.2. Many Small Clusters Are More Rigid but Cover Less Distance Variation. One might ask why a larger cluster number is less favorable. Obviously, the extreme case of defining each single pair of atoms as a separate cluster would lead to minimum STDDV within clusters but would represent a trivial and useless solution. Note that STTDV within clusters decreases with increasing N_{clust}, which makes clusters more homogeneous and is a desired effect. However, at the same time the overall amount of variability caught within clusters also decreases, which is an undesired effect, since larger portions of the molecule are disregarded. These trends are reflected by the total area of coloured squares along the diagonal in Figure 13. The smaller the squares (with increasing N_{clust}), the smaller their total area, even if there are more squares (the area decreases quadratic with the side lengths of squares).

This tendency has been quantitatively demonstrated for real MD-data in Figure 6. On top of that, Figure 14 shows that the same trend also holds for unstructured matrices, as mentioned above. However, results also show that matrices without structure (randomly rearranged elements) allow for very little reduction in the STDDV covered within clusters, as compared to structured matrices and that this difference further increases with N_{clust}.

4.3. Clustering Is Stable and Self-Contained. Powerful statistics on MD-trajectories (which is, e.g., able to detect small motions related to signaling) needs careful fitting of configuration frames as a prerequisite. Fitting to a domain which should be as rigid and large as possible is one of the options.

Hence, finding large clusters is desirable. However, large clusters in general might prove unstable. Therefore, we have carefully investigated this issue for the target function proposed by Bernhard and Noé [18] in conjunction with our clustering method.

It turned out that larger clusters are even more stable (see Figure 7), which is a strong indicator of stability and self-containment of our results.

4.4. Ergodicity. Atomic motions in a large molecule constitute a highly dimensional phase space, and MD-simulations can in most cases explore only part of the total phase space. At least, one can never be sure about the exact fraction of phase space actually visited in a specific simulation run. As a consequence we use only the second half of our MD trajectory for final clustering (see Figure 5), in accordance with our previous work [29].

4.5. Computational Resources. Clustering takes very little iteration steps for A_5, due to monotonous, continuous relationships between elements of **S**; see the appendix. Note that the rapidity of clustering in this case is not only a primary consequence of the small number of atoms but a matter of simplicity in structure.

Randomized matrices **S** take significantly more iterations for clustering, since many minima are almost equal regarding the target function, rendering solutions ambiguous. As opposed to this, clustering large molecules with structured internal motion, such as B4402 and B4403, yields well-defined minima after a reasonable number of trials.

4.6. Cluster Interpretation. There is an obvious difference in visual appearance between STDDV matrices for B4402 and B4403 as seen in Figure 4. Trajectory B4402 yields an unstructured, rather flat STDDV matrix (upper panel in Figure 4), while B4403 shows a distinctly structured STDDV matrix (lower panel in Figure 4). The relation between cluster rigidity and size is illustrated in Figure 7 and shows a clear trend: cluster rigidity increases with increasing cluster size, reflected in a decreasing STDDV within clusters. However, this does not mean that the largest cluster is always the most rigid cluster (i.e., has lowest STDDV). For $N_{\text{clust}} = 5$, we see that in B4403 the largest cluster is at the same time the most rigid one. For $N_{\text{clust}} = 5$ in B4402 the second largest cluster is

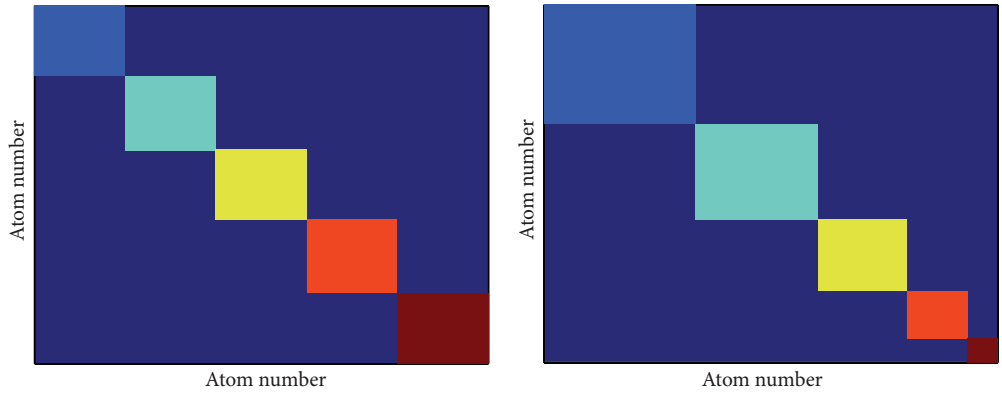


Figure 13: Clusters of equal and unequal sizes. If the pair-distance variations were randomly distributed, clusters of equal sizes were optimum (a). If clusters of unequal size result, this indicates that structure is present in the matrix (b). The figure is a schematic illustration for $N_{\text{clust}} = 6$.

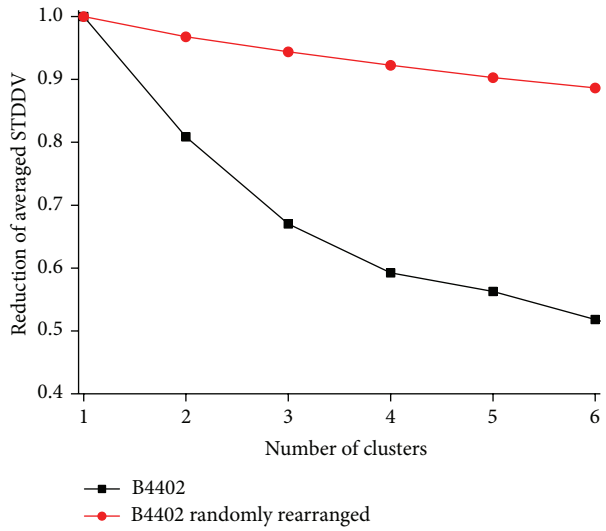


Figure 14: Improvement of target function with number of clusters. Comparison between data from MD trajectory B4402 and randomly rearranged matrix elements S_{ij}. Black graph "B4402" relates to matrix **S** obtained from a real MD-simulation, whereas the red graph "B4402 randomly rearranged" stems from a matrix with randomly rearranged elements (STDDVs).

the most rigid one. For $N_{\text{clust}} = 6$ the situation is inverted: in B4402 the largest cluster is the most rigid one. In B4403 the second largest cluster is most rigid.

All in all, the most rigid cluster was always found among the two largest clusters. The mappings of the clustering results for $N_{\text{clust}} = 5$ and $N_{\text{clust}} = 6$ for B4402 and B4403 have been displayed in Figure 5.

4.7. Conclusion and Prospects. A main result of the paper is the finding that the target function proposed by others [18] has only crisp solutions; that is, each atom belongs to one single cluster only (and not to several clusters in a fuzzy sense). This finding allows for a much more efficient search for optimum clustering.

Based on this new finding, the process of clustering was evaluated regarding various aspects to provide concomitant information for possible application by other investigators. Applicability was demonstrated for two trajectories of 250 ns each for large biomolecular complexes whose dynamics is of key importance for the understanding of immune reactions.

Further improvements can be expected from a more detailed investigation of the Kullback-Leibler distance [36, 37]. In this work it has only been reported as a means for assessing the quality of clustering by some given method. In future work, the Kullback-Leibler-distance may enter the clustering procedure itself and render clusters even more stable between subtrajectories and over time.

Appendices

A. Pair-Distance Variations in a Small Molecule

For comparison, we applied our clustering method also to the A_5 penta-L-alanine peptide already analyzed by Bernhard and Noé [18]. They choose penta-L-alanine to evaluate their clustering algorithm on MD simulations for reasons of simplicity of this molecule. A_5 has four amide bonds, each comprising four atoms (CONH). The delocalization of the nitrogen's free electron between conjugated carbonyl and amine groups poses planarity and rigidity onto this structure.

Mu et al. [38] showed that A_5 does not remain in an alpha-helical conformation, but rather exhibits repeated folding and unfolding events. As expected for such a molecule, our clustering of the averaged STDDV matrix indicates that atoms close together show little fluctuations of their mutual distances, such as the atoms in the middle of the pentapeptide; they are, so to speak, in the centre of the storm. Conversely, atoms near the edges show large distance fluctuations with respect to all other atoms.

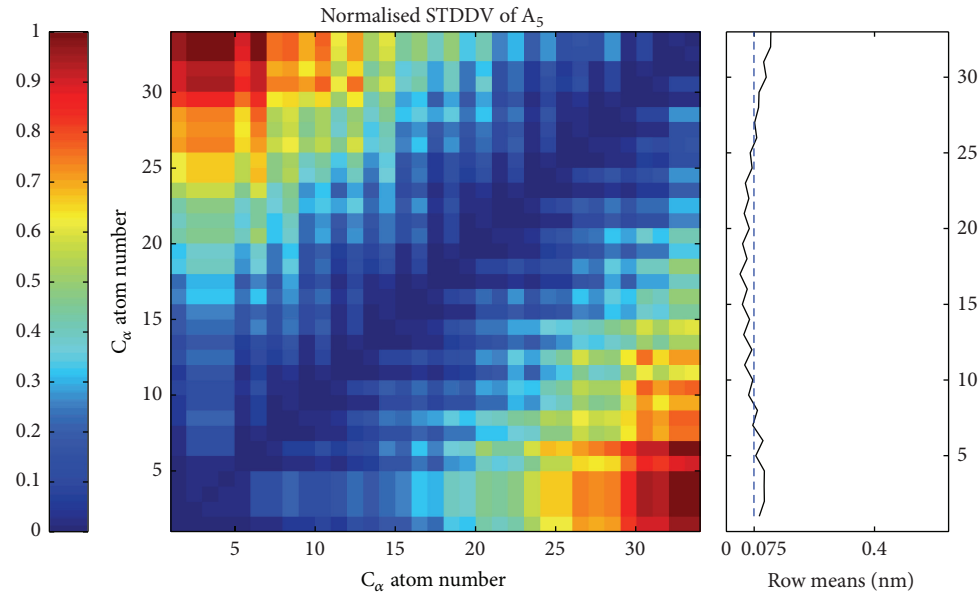


FIGURE 15: Matrix of standard deviation of distance variation (STDDV) for the A_5 pentapeptide (33 atoms). Values of STDDV [nm] are color coded (see bar on the left). Averaging over a row yields the mean distance variation against all other atoms (see subgraph on the right).

Clustering such a homogeneous matrix **S** does not reveal structural elements within the molecule, but rather splits it into equal parts, according to the number of clusters preselected. Note, however, that this behavior is a property of the equally distributed matrix values S_{ij} and not a general rule.

B. Software Availability

The software used for clustering into crisp domains of preselected number is currently being implemented in Java and will be made available for free download from http://www.meduniwien.ac.at/msi/biosim/index.php?lang=en&seite=en_forLehreIt_pairDistanceClustering.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The MD trajectories used in the present work were generated on the IBM-BlueGene computer facility at Bulgarian National Centre for Supercomputing Applications (NCSA). The work was supported in part by BSF and OeAD under Grants DCVP 02/1/2009, DNTS-A 01-2/2013, and WTA-BG 06/2013. We gratefully acknowledge mathematical advice and helpful discussions from Professor Rudolf Karch, PhD and Peicho Petkov, PhD, and technical assistance by Michael Cibena.

References

[1] G. Bao, "Mechanics of biomolecules," *Journal of the Mechanics and Physics of Solids*, vol. 50, no. 11, pp. 2237–2274, 2002.

[2] R. Lavery, A. Lebrun, J.-F. Allemand, D. Bensimon, and V. Croquette, "Structure and mechanics of single biomolecules: experiment and simulation," *Journal of Physics Condensed Matter*, vol. 14, no. 14, pp. R383–R414, 2002.

[3] D. Gordon, R. Chen, and S. H. Chung, "Computational methods of studying the binding of toxins from venomous animals to biological ion channels: theory and applications," *Physiological Reviews*, vol. 93, pp. 767–802, 2013.

[4] Y. Cui, "Using molecular simulations to probe pharmaceutical materials," *Journal of Pharmaceutical Sciences*, vol. 100, no. 6, pp. 2000–2019, 2011.

[5] S. A. Adcock and J. A. McCammon, "Molecular dynamics: survey of methods for simulating the activity of proteins," *Chemical Reviews*, vol. 106, no. 5, pp. 1589–1615, 2006.

[6] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[7] W. Wriggers and K. Schulten, "Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates," *Proteins*, vol. 29, pp. 1–14, 1997.

[8] W. R. Taylor, "Protein structural domain identification," *Protein Engineering*, vol. 12, no. 3, pp. 203–216, 1999.

[9] J. M. Haile, "Fundamentals," in *Molecular Dynamics Simulation: Elementary Methods*, John Wiley & Sons, 1992.

[10] J. M. Haile, "Hard spheres," in *Molecular Dynamics Simulation: Elementary Methods*, John Wiley & Sons, 1992.

[11] H. J. Berendsen and S. Hayward, "Collective protein dynamics in relation to function," *Current Opinion in Structural Biology*, vol. 10, no. 2, pp. 165–169, 2000.

[12] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," *Proteins: Structure, Function and Genetics*, vol. 17, no. 4, pp. 412–425, 1993.

[13] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten, "Principal component analysis and long time protein dynamics," *Journal of Physical Chemistry*, vol. 100, no. 7, pp. 2567–2572, 1996.

[14] S. Hayward and B. L. De Groot, "Normal modes and essential dynamics," *Methods in Molecular Biology*, vol. 443, pp. 89–106, 2008.

[15] A. E. García, "Large-amplitude nonlinear motions in proteins," *Physical Review Letters*, vol. 68, no. 17, pp. 2696–2699, 1992.

[16] X. Zhang, D. Bhatt, and D. M. Zuckerman, "Automated sampling assessment for molecular simulations using the effective sample size," *Journal of Chemical Theory and Computation*, vol. 6, no. 10, pp. 3048–3057, 2010.

[17] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, "GRGMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008.

[18] S. Bernhard and F. Noé, "Optimal identification of semi-rigid domains in macromolecules from molecular dynamics Simulation," *PLoS ONE*, vol. 5, no. 5, Article ID e10491, 2010.

[19] J. Robinson, J. A. Halliwell, H. McWilliam, R. Lopez, P. Parham, and S. G. Marsh, "The IMGT/HLA database," *Nucleic Acids Research*, vol. 41, pp. D1222–D1227, 2013.

[20] W. A. Macdonald, Z. Chen, S. Gras et al., "T cell allorecognition via molecular mimicry," *Immunity*, vol. 31, no. 6, pp. 897–908, 2009.

[21] N. Guex and M. C. Peitsch, "SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling," *Electrophoresis*, vol. 18, no. 15, pp. 2714–2723, 1997.

[22] H. J. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, and J. Hermans, "Interaction models for water in relation to protein hydration," *Intermolecular Forces*, vol. 14, pp. 331–342, 1981.

[23] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.

[24] U. Omasits, B. Knapp, M. Neumann et al., "Analysis of key parameters for molecular dynamics of pMHC molecules," *Molecular Simulation*, vol. 34, no. 8, pp. 781–793, 2008.

[25] S. O. Yesylevskyy, V. N. Kharkyanen, and A. P. Demchenko, "Hierarchical clustering of the correlation patterns: new method of domain identification in proteins," *Biophysical Chemistry*, vol. 119, no. 1, pp. 84–93, 2006.

[26] T. Shibuya, "Fast hinge detection algorithms for flexible protein structures," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 2, pp. 333–341, 2010.

[27] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[28] M. P. Allen, *Introduction to Molecular Dynamics Simulation*, John von Neumann Institute for Computing, 2004.

[29] W. Schreiner, R. Karch, B. Knapp, and N. Ilieva, "Relaxation estimation of RMSD in molecular dynamics immuno-simulations," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 173521, 9 pages, 2012.

[30] J. B. Clarage, T. Romo, B. K. Andrews, B. M. Pettitt, and G. N. Phillips Jr., "A sampling problem in molecular dynamics simulations of macromolecules," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 8, pp. 3288–3292, 1995.

[31] V. Daggett, "Long timescale simulations," *Current Opinion in Structural Biology*, vol. 10, no. 2, pp. 160–164, 2000.

[32] B. Hess, "Convergence of sampling in protein simulations," *Physical Review E. Statistical, Nonlinear, and Soft Matter Physics*, vol. 65, no. 3, Article ID 031910, 2002.

[33] L. J. Smith, X. Daura, and W. F. Van Gunsteren, "Assessing equilibration and convergence in biomolecular simulations," *Proteins. Structure, Function and Genetics*, vol. 48, no. 3, pp. 487–496, 2002.

[34] A. Grossfield and D. M. Zuckerman, "Quantifying uncertainty and sampling quality in biomolecular simulations," *Annual Reports in Computational Chemistry*, vol. 5, pp. 23–48, 2009.

[35] H. Flyvbjerg and H. G. Petersen, "Error estimates on averages of correlated data," *The Journal of Chemical Physics*, vol. 91, no. 1, pp. 461–466, 1989.

[36] C. L. McClendon, L. Hua, A. Barreiro, and M. P. Jacobson, "Comparing conformational ensembles using the Kullback-Leibler divergence expansion," *Journal of Chemical Theory and Computation*, vol. 8, pp. 2115–2126, 2012.

[37] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[38] Y. Mu, P. H. Nguyen, and G. Stock, "Energy landscape of a small peptide revealed by dihedral angle principal component analysis," *Proteins. Structure, Function and Genetics*, vol. 58, no. 1, pp. 45–52, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 472869, 9 pages
http://dx.doi.org/10.1155/2014/472869

Research Article

Computational Study to Determine When to Initiate and Alternate Therapy in HIV Infection

Matthias Haering,[1] Andreas Hördt,[2] Michael Meyer-Hermann,[1,3] and Esteban A. Hernandez-Vargas[1]

[1] *Department of Systems Immunology and Braunschweig Integrated Centre of Systems Biology, Helmholtz Centre for Infection Research, Inhoffenstraße 7, 38124 Braunschweig, Germany*
[2] *Institut für Geophysik und extraterrestrische Physik, Technische Universität Braunschweig, Mendelssohnstraße 3, 38106 Braunschweig, Germany*
[3] *Institute for Biochemistry, Biotechnology and Bioinformatics, Braunschweig University of Technology, Braunschweig, 38106 Braunschweig, Germany*

Correspondence should be addressed to Esteban A. Hernandez-Vargas; esteban.vargas@theoretical-biology.de

Received 15 February 2014; Revised 7 April 2014; Accepted 10 April 2014; Published 11 May 2014

Academic Editor: Filippo Castiglione

HIV is a widespread viral infection without cure. Drug treatment has transformed HIV disease into a treatable long-term infection. However, the appearance of mutations within the viral genome reduces the susceptibility of HIV to drugs. Therefore, a key goal is to extend the time until patients exhibit resistance to all existing drugs. Current HIV treatment guidelines seem poorly supported as practitioners have not achieved a consensus on the optimal time to initiate and to switch antiretroviral treatments. We contribute to this discussion with predictions derived from a mathematical model of HIV dynamics. Our results indicate that early therapy initiation (within 2 years postinfection) is critical to delay AIDS progression. For patients who have not received any therapy during the first 3 years postinfection, switch in response to virological failure may outperform proactive switching strategies. In case that proactive switching is opted, the switching time between therapies should not be larger than 100 days. Further clinical trials are needed to either confirm or falsify these predictions.

1. Introduction

According to the last report from UNAIDS in 2011 [1], 34 million people live with human immunodeficiency virus (HIV). Although the number of new cases of HIV infection is declining, the number of people living with HIV is increasing; therefore the problems of continuing treatment of chronic infection are of major importance in today's social medicine.

Nowadays, the drugs to treat HIV type 1 (HIV-1) infection belong to four distinct classes [2, 3]: reverse transcriptase inhibitors, protease inhibitors, integrase inhibitors, and fusion inhibitors (Figure 1). Currently, highly active antiretroviral therapies (HAART) generally comprise three different drugs to combat different parts of the HIV cycle. These therapies prevent immune deterioration, reduce morbidity and mortality, and prolong the life expectancy of people infected with HIV.

Nevertheless, HAART is not always successful. Many patients have long-term complications while others experience virological failure (inability to maintain HIV RNA levels below 50 copies/mL) [3]. In most cases, viral rebound is associated with the emergence of resistance-conferring mutations within the viral genome, resulting in reduced viral susceptibility to one or more of the drugs. This is related to the reverse transcription process of viral RNA into DNA, which is highly prone to errors, introducing on average one mutation for each viral genome transcribed [2].

The primary goal of the initial regimen proposed in the guidelines for the use of antiretroviral agents in HIV-1 infected adults and adolescents by Department of Health and

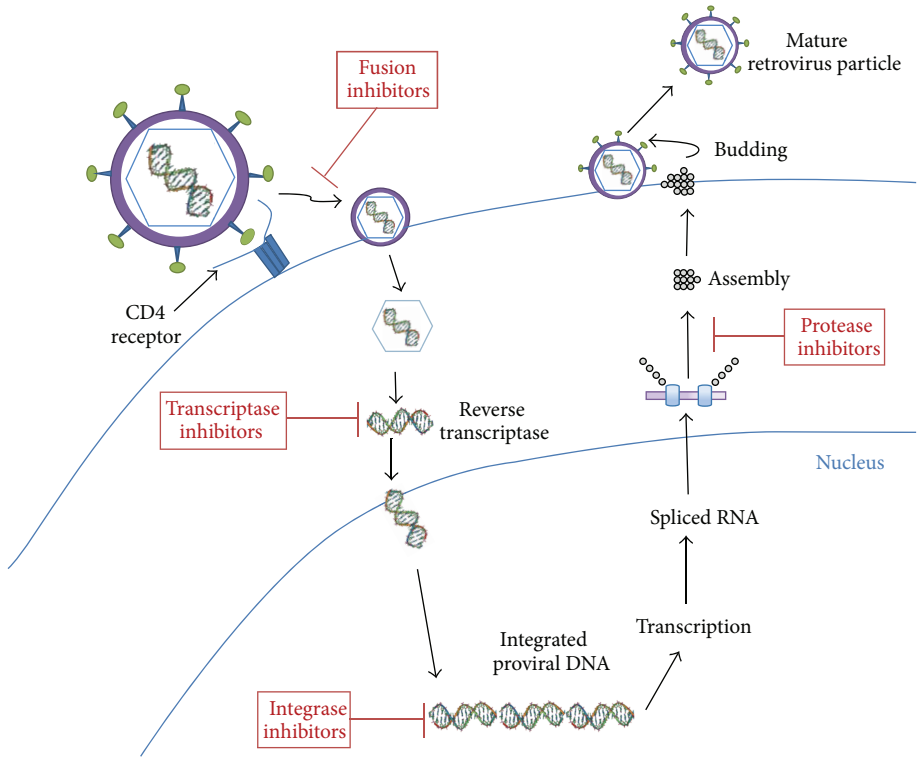


Figure 1: HIV infection cycle affected by the four distinct drug classes. The drug classes are shown in red boxes.

Human Services (DHHS) [3] is to suppress viral replication to the maximum and to sustain this level of suppression as long as possible. Even though HAART can reduce the viral load in the blood by at least five orders of magnitude, ongoing low-level replication still occurs; hence, the risk of developing resistance is always present. Complete virus eradication by HAART does not appear to be achievable in the foreseeable future. Furthermore, antiretroviral guidelines [3] have not achieved a consensus on two fundamental questions in HIV treatment: (i) when to start antiretroviral treatment and (ii) when to change for a new antiretroviral treatment.

The standard procedure for therapy initiation is still a point of discussion in the expert panel over the last 20 years [3]. However, there is a general consensus that antiretroviral therapy should be initiated in all patients with a history of an AIDS-defining illness or when CD4+ T cell counts are less than 350 cells/mm^3. It is strongly recommended to start therapy if the CD4+ T cell count is between 350 and 500 cells/mm^3 and there is a recommendation with moderate urgency for patients with CD4+ T cell counts > 500 cells/mm^3 (Table 1).

The expert panel [3] pointed out the absence of cohort studies that conclusively demonstrate a clinical benefit of HAART in patients with CD4+ T count > 350 cells/mm^3. For some patients, the potential risks of short- or long-term drug-related complications and nonadherence to long-term therapy may offset possible benefits of earlier therapy initiation.

Computational biology may play an important role in evaluating the impact of the initiation time of HAART during HIV infection. Using a mathematical model, [4] suggested to initiate treatment in stages of the infection when the viral load can be easily controlled: in the acute phase of the infection when the viral load peaks and moderately in the asymptomatic phase. The very early phase and the AIDS phase are considered hardly controllable. Thus, authors in [4] argued in favour of an early but not immediate treatment initiation. The main drawback of this study is the lack of a mathematical model that is able to reproduce the whole disease trajectory, which limits the long-term assessments of treatment strategies. The observation that HAART timing has a strong impact on the disease outcome is supported by computational results by [5, 6]. However, these studies have similar drawbacks, as they do not envisage long-term dynamics (more than 8 years when AIDS may appear) and different treatment protocols.

Table 1: Panel recommendations for therapy initiation.

		Urgency of initiation
CD4+ T cell count	<350 cells/mm^3	Strong I
	350–500 cells/mm^3	Strong II
	>500 cells/mm^3	Moderate III
Transmission risk	Perinatal	Strong I
	Heterosexual	Strong I
	Other	Strong III

The urgency is rated by validity: I: data from randomized trials, II: data from well-designed nonrandomized trials or observational cohort studies with long-term clinical outcomes, III: expert opinion [3].

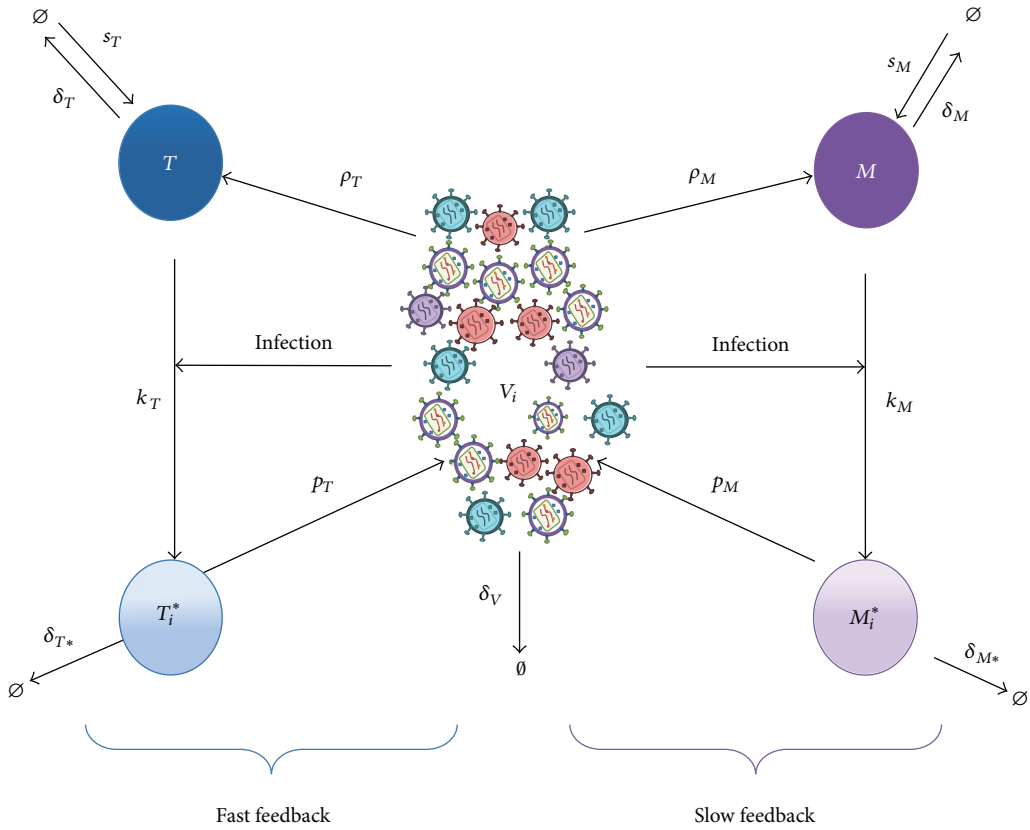


FIGURE 2: Nonlinear HIV model. T represents the uninfected CD4+ T cells, T_i^* represents the infected CD4+ T cells with the ith strain, M represents uninfected macrophages, M_i^* represents infected macrophages with the ith strain, V_i represents the ith strain, and V_T is the sum of all strains.

For the second fundamental question in HIV treatment, when treatment should be alternated in response to virological failure, the answer is more complex [3, 7, 8]. One clinical goal is to delay the time until patients exhibit strains resistant to all existing regimens. There is a crucial trade-off between switching therapies. On the one hand, switching early carries the risk of poor adherence to a new drug regimen and prematurely exhausting the limited number of remaining salvage therapies. On the other hand, switching drugs too late allows the accumulation of mutations that leads to multidrug resistance [2, 3]. The most aggressive approach would be to change therapies for any repeated detectable viremia (e.g., two consecutive HIV RNA > 50 copies/mL after suppression). The most conservative strategy has been to allow detectable viremia up to an arbitrary level (e.g., 1000–500 copies/mL). This latter approach is called switch on virological failure [3].

Recently, Martinez-Picado et al. [9] suggested that switching between therapies can decrease the likelihood of viral resistance and prolong the pre-AIDS period. This hypothesis was supported by clinical trials [9] called *Switching Antiviral Therapy Combination against HIV* (SWATCH), which consists of two HAART regimens that are periodically alternated every 3 months. The alternation of 2 drug regimens could inhibit the emergence of highly resistant genotypes.

In this paper, we focus on the trade-off between the inhibition of long-term reservoirs and the promotion of resistant genotypes and thus help to precisely work out the advantages of early treatment. We also investigate how frequently antiretroviral regimens should be alternated in order to attain optimal proactive treatments.

2. Material and Methods

2.1. Mathematical Model. A typical HIV infection response consists of three stages: an initial acute infection, a long asymptomatic period, and a final increase in viral load with simultaneous collapse in healthy CD4+ T cell counts. The majority of existing mathematical models [4, 11–20] give a good representation of either the first two stages [11, 13, 21–24] or the last stage of the infection [25] but do not describe the three stages observed in HIV infection in a single framework. A mathematical model that is able to represent the typical HIV infection dynamics including all three stages was suggested by [26]. The model includes the possibility of full parameter variations, without losing the capability to describe the three stages.

The results in [26] indicate that HIV infection can be considered as two feedback systems (Figure 2). One provides the fast dynamics presented in the early stages of infection as a result of the fast infection process of CD4+ T cells. The second feedback sustains a constant slow infection process in

macrophages over the years accompanied by the long time survival conditions of macrophages.

Here, following the work by [8], we extend the mutation tree from [10] into a nonlinear model with mutations (1)–(5). Unlike existing models [4, 8, 11–20], our model is able to adequately represent the three stages of HIV infection and the dynamics of resistant genotypes when HAART treatment is introduced. The model is defined by the following set of differential equations:

$$\dot{T} = s_T + \frac{\rho_T}{C_T + V_T} T V_T - \sum_{i=1}^{n} k_T f_i \left(1 - \eta_{\sigma,i}^T\right) T V_i - \delta_T T, \quad (1)$$

$$\dot{M} = s_M + \frac{\rho_M}{C_M + V_T} M V_T - \sum_{i=1}^{n} k_M f_i \left(1 - \eta_{\sigma,i}^M\right) M V_i - \delta_M M, \quad (2)$$

$$\dot{T}_i^* = k_T f_i \left(1 - \eta_{\sigma,i}^T\right) T V_i + \sum_{j=1}^{n} \mu m_{i,j} V_j T - \delta_{T^*} T_i^*, \quad (3)$$

$$\dot{M}_i^* = k_M f_i \left(1 - \eta_{\sigma,i}^M\right) M V_i + \sum_{j=1}^{n} \mu m_{i,j} V_j M - \delta_{M^*} M_i^*, \quad (4)$$

$$\dot{V}_i = p_T f_i \left(1 - \theta_{\sigma,i}^T\right) T_i^* + p_M f_i \left(1 - \theta_{\sigma,i}^M\right) M_i^* - \delta_V V_i, \quad (5)$$

where T represents the uninfected CD4+ T cells, T_i^* represents the infected CD4+ T cells with the ith strain (genetic variant or subtype of the virus), M represents uninfected macrophages, M_i^* represents infected macrophages with the ith strain, V_i represents the ith strain, and V_T is the sum of all n strains.

Parameters s_T and s_M represent the source terms of new CD4+ T cells and macrophages, respectively. HIV, as other pathogens, triggers the proliferation of immune cells. Homeostatic proliferation is modeled using a logistic growth model limited by viral load. This would allow convergence to a high percentage of the reservoirs (macrophages) being infected without allowing the total population to expand at unrealistic growth rates. Parameters ρ_T and ρ_M are the maximum proliferation rate for CD4+ T cells and macrophages, respectively. C_T and C_M represent the respective half-velocity constants.

The infection rate constant is represented with k_T for CD4+ T cells and k_M for macrophages. Viral proliferation is achieved in infected CD4+ T cells and infected macrophages with rate constants p_T and p_M, respectively. These parameters depend on the fitness of the genotype (f_i) and the therapy (σ) that is being used.

We consider two therapies ($\sigma = 1, 2$) composed of reverse transcriptase inhibitors with effectiveness $\eta_{\sigma,i}^T$ for CD4+ T cells and $\eta_{\sigma,i}^M$ for macrophages and protease inhibitors with effectiveness $\theta_{\sigma,i}^T$ for CD4+ T cells and $\theta_{\sigma,i}^M$ for macrophages. Note that, based on clinical evidence [27], inhibitors are more effective in CD4+ T cells than in macrophages; this is considered by $\eta_{\sigma,i}^T > \eta_{\sigma,i}^M$ and $\theta_{\sigma,i}^T > \theta_{\sigma,i}^M$. The therapy parameter values used in this study are given in Figure 3.

The mutation rate is expressed by μ, and $m_{i,j} \in \{0, 1\}$ represents the genetic connections between genotypes.

Table 2: Parameter values for (1)–(5).

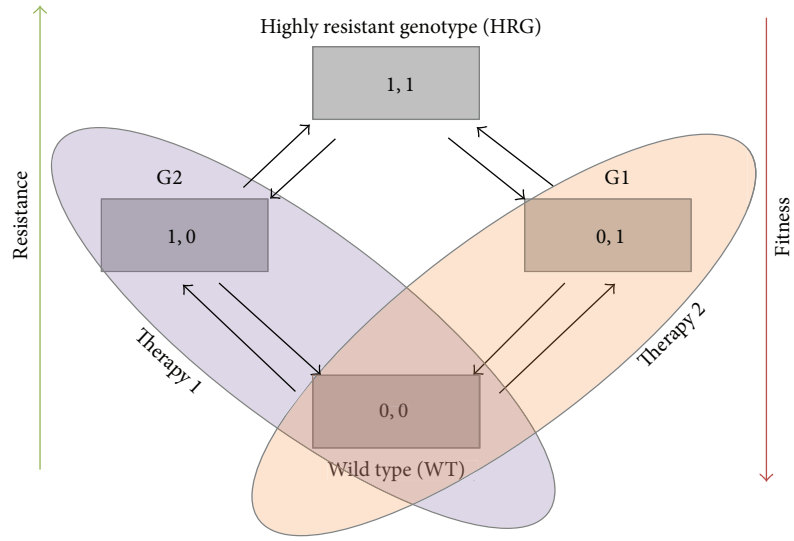
Parameter	Units	Nominal value	Source
s_T	Cells/mm^3 day	10	[28]
s_M	Cells/mm^3 day	0.15	[28]
k_T	mm^3/day copies	4.57×10^{-5}	[28]
k_M	mm^3/day copies	4.33×10^{-8}	[26]
p_T	Copies/cell day	38	[29]
p_M	Copies/cell day	35	[29]
δ_T	Day^{-1}	0.01	[29]
δ_T^*	Day^{-1}	0.4	[29]
δ_M	Day^{-1}	1×10^{-3}	[26]
δ_M^*	Day^{-1}	1×10^{-3}	[26]
δ_V	Day^{-1}	2.4	[29]
ρ_T	Day^{-1}	0.01	[26]
ρ_M	Day^{-1}	0.003	[26]
C_T	Copies/mm^3	300	[26]
C_M	Copies/mm^3	220	[26]

The degradation rates for the relevant species are δ_T, δ_{T^*}, δ_M, δ_{M^*}, and δ_V. Parameter values are presented in Table 2.

The meaning of the different terms and parameters is illustrated in a block diagram (Figure 2). For example, the left-hand side of (1) defines the rate of change of the uninfected CD4+ T cells. The right-hand side defines how CD4+ T cells are altered during infection: a constant production of CD4+ T cells is expressed by s_T. CD4+ T cell depletion is proportional to the death rate δ_T and the current number of healthy CD4+ T cells. The cell proliferation is proportional to the number of healthy cells and the total viral load V_T, which represents the sum over all strains (n). The infection is described by the sum term on the right-hand side of (1) and is also proportional to the existing number of uninfected cells and the strains V_i. Each strain has its own infection rate constant $k_T f_i(1 - \eta_{\sigma,i}^T)$, which is why we have to sum over index i. The other equations may be understood in a similar fashion.

2.2. Viral Mutation Tree. Nowadays, it is considered critical to take viral mutation into account during the development of treatment strategies. The process of reverse transcription is extremely error-prone and it is during this step that mutations can occur. High levels of resistance can be produced by substitutions of a single amino acid [3]. For instance, when lamivudine is used as a single agent, resistant strains will appear in a few weeks [2]. This is the reason why monotherapy has been discontinued and HAART is composed of at least three different drugs. As a result, multiple mutations are required for resistance to occur to all drugs in one regimen.

As a simple motivating example, we consider the mutation tree with 4 variants and 2 possible antiretroviral treatments proposed in [10]. The wild type genotype (WT) would be the most prolific variant in the absence of any drugs (Figure 3). However, it is also the variant that all drug combinations have been designed to combat and therefore is susceptible to all therapies. The wild type genotype (WT) can mutate to either genotype 1 ($G1$) which is susceptible to therapy 2 or genotype 2 ($G2$) which is susceptible to therapy 1. After mutations, the highly resistant genotype (HRG) is



Genotype (i)		WT	G1	G2	HRG
Fitness (f_i)		1.0	0.8	0.8	0.7
Therapy 1 $\sigma = 1$	CD4+ T cells ($\theta^T_{\sigma,i} = \eta^T_{\sigma,i}$)	0.5	0.01	0.3	0.01
	Macrophages ($\theta^M_{\sigma,i} = \eta^M_{\sigma,i}$)	0.3	0.01	0.1	0.01
Therapy 2 $\sigma = 2$	CD4+ T cells ($\theta^T_{\sigma,i} = \eta^T_{\sigma,i}$)	0.5	0.3	0.01	0.01
	Macrophages ($\theta^M_{\sigma,i} = \eta^M_{\sigma,i}$)	0.3	0.1	0.01	0.01

Figure 3: Four variant mutation trees. The wild type (*WT*) is susceptible to both therapies. Genotype 1 (*G1*) is susceptible to therapy 2 and genotype 2 (*G2*) is susceptible to therapy 1. The highly resistant genotype (*HRG*) is not affected by any therapy. Parameter values were taken from [10].

a genotype with a low proliferation rate, but resistant to all drug therapies.

2.3. Therapy. The DHHS panel [3] recommends therapies with two nucleoside analogues and either protease inhibitors or nucleoside reverse transcriptase inhibitors. The combination of these drugs is crucial in controlling the development of resistance. For simulation purposes, we consider the common clinical strategy suggested by the DHHS panel in [3], which recommends change therapy when virological failure is presented, that is, switch on virological failure (SVF): introduce a new regimen if there is detectable viremia (HIV RNA > 1,000 copies/mL) and drug-resistant genotypes are identified. We also consider the SWATCH strategy proposed in [9]. The rationale behind this strategy is that one could preempt virologic rebound and reduce accumulating drug-resistant genotypes by alternating treatments. This strategy is implemented as follows:

> *SWATCH*: alternate between two regimens every 3 months.

2.4. Monte Carlo Simulations. As the interpatient variability and the extreme error sensitivity of the HIV replication process are nonnegligible in HIV treatment, we perform Monte Carlo (MC) simulations to analyse treatment strategies in stochastic environments [30]. Our MC algorithm consists of repeated random sampling, providing numerical results that can be interpreted with statistical methods. To this end, we consider 1000 repeated simulations with randomly perturbed parameters (normal distribution and 30% deviation from the nominal parameter values in Table 2). The described MC simulations without therapy, with SVF, and SWATCH treatments were carried out at six different initiation times t_i (0.5, 1, 1.5, 2, 3, and 4 years). The differential equations (1)–(5) are solved with the toolbox *ode*45 from the MATLAB library. We record the year when immunological failure appears (t_f), that is, when CD4+ T cells sink under 200 copies/mm^3. N_p represents the number of cases where no immunological failure occurs during 30 years. The MC procedure is illustrated as a flow diagram in Figure 4. The analysis of the MC simulation results is based on a two-way ANOVA test. The significance is identified as P value < 0.05; the data are further analyzed by a two-way t-test and a Bonferroni posttest.

3. Results and Discussion

For the scenario with no treatment, we obtain an average time of immunological failure of 8.5 years postinfection from our MC simulations. In addition, 99% of the simulated cases may

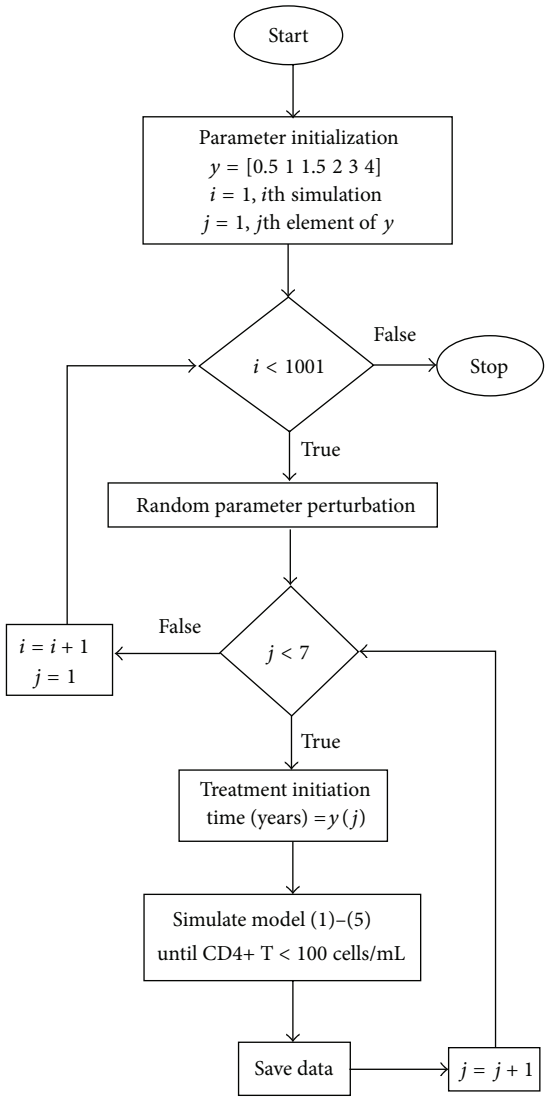


Figure 4: Monte Carlo simulation algorithm. The Simulation box includes the simulation of the four-genotype mutation model under a certain treatment. For each of the six times of treatment initiation, the simulation will be repeated 1000 times and the data can be analysed.

progress to AIDS at different time scales. These results are consistent with clinical observations in [31], suggesting that model (1)–(5) can adequately represent the basic features of HIV infection.

3.1. Treatment Initiation. The treatment strategies SVF and SWATCH (Figure 5) were implemented in model (1)–(5) using different initiation times t_i (0.5, 1, 1.5, 2, 3, and 4 years) and parameter values as presented in Table 2. During the first 5 years, we observe that both SVF and SWATCH satisfy the levels required for healthy immunological responses (CD4+ T > 500 cells/mm^3) and decrease the viral load to undetectable levels (<50 copies/mL) as suggested in [3].

For both treatment strategies, our results support the hypothesis that an early treatment initiation is more beneficial than a late start of the therapy. However, simulation results for the SVF strategy (Figure 5) suggest that there is only little improvement for a very early treatment initiation (<1 year).

If a second treatment is introduced, a second virological failure may appear, faster than the first one, progressing to AIDS (Figure 5). This is consistent with the observation in [32] that persistent low-level viremia and long-term reservoirs promote a second virological failure.

The clinical trial SWATCH [9] suggested an alternative solution to minimize HIV resistance mutation by alternating between two regimens every three months while viral load is suppressed. In our simulations, the alternation is visible in form of the high-frequency oscillations. For early initiation treatment (<2 years), the results in Figure 5 indicate that the SWATCH approach is clearly superior to SVF because the time for immunological failure may occur 10 years later compared to the SVF strategy. However, SWATCH is more sensitive to the initiation time. For instance, when $t_i = 4$ years, SWATCH and SVF may show similar performance. We conclude that SWATCH treatment may provide significant extension of the time to virological failure only if the treatment is initiated before the third year postinfection.

While the results above were obtained for one particular parameter set, we now consider the nonlinearity of the problem and the corresponding sensitivity to parameter variations by using an MC approach with 1000 random samples (Table 3). In comparison to the MC simulations without treatment, we note in Table 3 that the number of cases without immunological failure (N_p) increases substantially (approximately 30%). For instance, when $t_i = 0.5$, MC simulations suggest that SVF may achieve 29.8% cases without immunological failure while SWATCH may achieve up to 46.6%. Furthermore, the appearance of virological failure is prolonged approximately by 4–7 years (P value ≤ 0.05).

Our studies support the hypothesis that an early intervention has significant positive impact on postponing the progression to AIDS. Figure 6 shows the average time when immunological failure occurs ($\bar{t}_f$) as a function of the initial time (t_i). We can see that SVF performance decreases almost linearly with respect to the initial time, while SWATCH shows approximately a parabolic behaviour. Note that the SWATCH strategy can outperform the SVF strategy only when therapy is initiated before the second year postinfection (P value ≤ 0.05). Therefore, it is not recommended to use the SWATCH strategy on patients who have not initiated treatment during 3 years postinfection.

Our results lead us to recommend avoiding treatment initiation after 2 years postinfection. In addition, for patients who have not received any treatment within 3 years or more postinfection, the SVF strategy is a better alternative since the advantage of SWATCH is fading (Figure 6) while the risk of long-term drug toxication could be smaller with the SVF strategy.

3.2. Alternating between Treatments. Computational studies [7, 8, 10] and clinical trials [9] suggest that a proactive alternating strategy like SWATCH may yield promising results. However, not enough work has been done to analyse

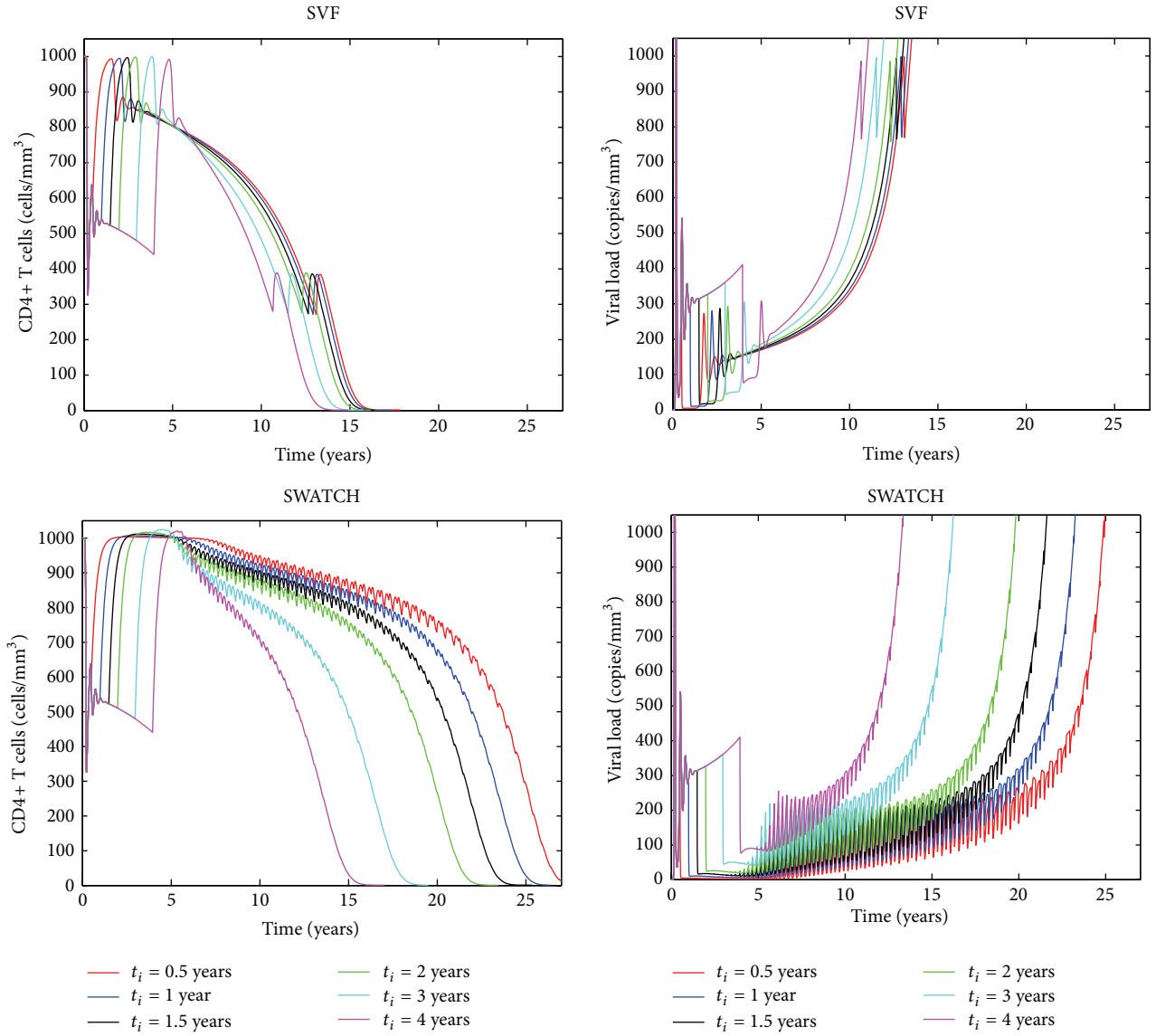


FIGURE 5: SVF and SWATCH treatment strategies. CD4+ T cells and viral load are displayed for different initiation times (t_i). The top two panels show simulation results for the SVF strategy and the bottom panels show the results for the SWATCH strategy. Deterministic simulations are based on nominal parameter values from Table 2.

the synergistic effect of initiation time and period of alternation between treatments.

Here, we ran model (1)–(5) with different switching and different initiation times. The first observation is that fast switching times between regimens yield longer periods without immunological failure (Figure 7). In addition, for switching times larger than 150 days, simulation trajectories reached a steady state, meaning that long periods between treatment switches should be avoided. An aspect that is not covered here, however, is that fast switching between regimens (<60 days) could increase drug toxicity and may lead to bad adherence to therapy. Further studies (pharmacokinetics/pharmacodynamics) are needed to include such issues to develop a comprehensive management concept.

A second important conclusion from Figure 7 is that the SWATCH strategy seems to perform poorly if the treatment was initiated 2 years postinfection. To attain a high effectiveness of the treatment, the switching frequency between antiretroviral regimens may need to be increased depending on the delay of therapy initiation. Considering $t_i = 3$ or $t_i = 4$ years, our results suggest that very short switching times about 10–60 days would be required to obtain the same performance as if the therapy was started earlier. Consistent with the results in Figure 6, this investigation leads to the conclusion that the SWATCH strategy should be initiated within 2 years postinfection to achieve responses superior to common clinical approaches (SVF).

Our results reveal the importance of early therapy to delay the AIDS progression. There are hints that an early intervention could improve the patient healing process. The recent clinical results by [33] showed the possibility of a mechanistic cure for patients who are treated immediately

Table 3: Monte Carlo simulation outcome for SVF and SWATCH strategies.

(a) SVF

t_i (years)	$\bar{t}_f \pm \bar{\sigma}$ (years)	N_p
0.5	14.01 ± 4.34	298
1	13.93 ± 4.39	292
1.5	13.82 ± 4.43	284
2	13.61 ± 4.38	277
3*	13.06 ± 4.08	259
4*	12.83 ± 4.55	253

(b) SWATCH

t_i (years)	$\bar{t}_f \pm \bar{\sigma}$ (years)	N_p
0.5	15.54 ± 4.79	466
1	15.45 ± 4.81	454
1.5	15.13 ± 4.79	442
2*	14.69 ± 4.68	427
3*	13.67 ± 4.38	386
4*	12.59 ± 3.84	336

$\bar{t}_f$ is the average year when immunological failure occurs for the 1000 random samples and the respective standard deviation $\bar{\sigma}$. N_p is the number of cases that do not experience immunological failure after 30 years postinfection. *Star represents a statistically significant difference with respect to the group of $t_i = 0.5$ years (P-value ≤ 0.05). There is a significant difference between SWATCH and SVF for all the groups when $t_i \leq 2$.

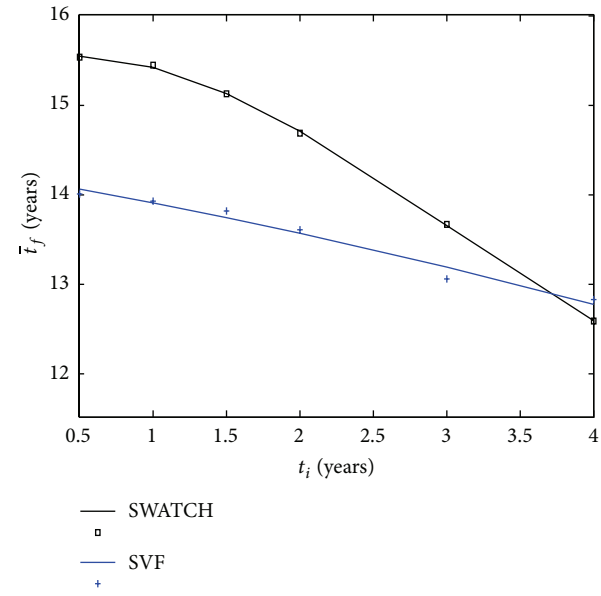


Figure 6: Monte Carlo simulation outcome for SVF (+) and SWATCH (□) strategies. Averaged time of immunological failure ($\bar{t}_f$) based on 1000 random samples is plotted against different initiation times (t_i). A parabolic function was fitted to the SWATCH data (solid black line) and a linear function to the SVF data (solid blue line).

after the infection. A mechanistic cure means a permanent viral suppression in the absence of therapy to levels that prevent immunodeficiency and transmission. However, it was not possible to simulate this behaviour using model (1)–(5). This is likely because the model is based on ODEs, implying that populations arbitrarily close to zero can recover. Discrete approaches could bring new insights into this field.

4. Conclusions

From our studies, we conclude that antiretroviral treatment strategies initiated after 2 years postinfection are not beneficial to extend the time to progression to AIDS.

Another significant result is that the SWATCH strategy outperforms SVF only when therapies are initiated within 2 years postinfection and switching periods for SWATCH strategy are less than 90 days. Large switching periods between regimens (>100 days) should be avoided during the application of SWATCH.

This work is a step forward for defining criteria for when to initiate and alternate therapy. Future work will be directed to the experimental evaluation of the presented results.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

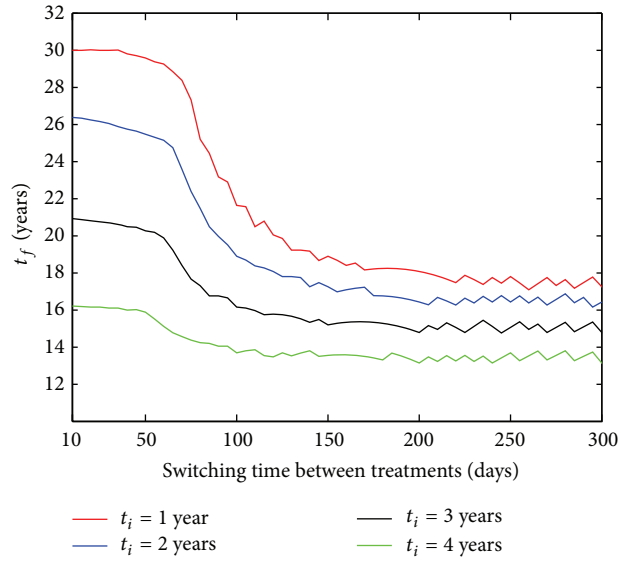


Figure 7: Optimal switching between regimens. Time of immunological failure (t_f) with respect to switching time for different initiation times (t_i).

Acknowledgments

This work was supported by BMBF-GerontoSys initiative (GerontoShield), portfolio HGF (metabolic dysfunction), and BMBF e-Med project SYSIMIT.

References

[1] UNAIDS, *Global Report: UNAIDS Report on the Global AIDS Epidemic 2010*, 2010.

[2] F. Clavel and A. J. Hance, "HIV Drug Resistance," *The New England Journal of Medicine*, vol. 350, no. 10, pp. 1023–1035, 2004.

[3] DHHS, *Panel on Antiretroviral Guidelines for Adults and Ado-Lescents. Guidelines for the Use of Antiretroviral Agents in HIV-1-Infected Adults and Adolescents*, Department of Health and Human Services (DHHS), 2013.

[4] A. M. Jeffrey, X. Xia, and I. K. Craig, "When to initiate HIV therapy: a control theoretic approach," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 11, pp. 1213–1220, 2003.

[5] F. Castiglione and P. Paci, "Criticality of timing for anti-HIV therapy initiation," *PLoS ONE*, vol. 5, no. 12, Article ID e15294, 2010.

[6] P. Paci, F. Martini, M. Bernaschi, G. D'Offizi, and F. Castiglione, "Timely HAART initiation may pave the way for a better viral control," *BMC Infectious Diseases*, vol. 11, article 56, 2011.

[7] E. A. Hernandez-Vargas, P. Colaneri, and R. H. Middleton, "Optimal therapy scheduling for a simplified HIV infection model," *Automatica*, vol. 49, pp. 2874–2880, 2013.

[8] E. A. Hernandez-Vargas, P. Colaneri, and R. H. Middleton, "Switching strategies to mitigate HIV mutation," *IEEE Transactions on Control Systems Technology*, 2014.

[9] J. Martinez-Picado, E. Negredo, L. Ruiz et al., "Alternation of antiretroviral drug regimens for HIV infection: a randomized, controlled trial," *Annals of Internal Medicine*, vol. 139, no. 2, pp. 81–89, 2003.

[10] E. Hernandez-Vargas, P. Colaneri, R. Middleton, and F. Blanchini, "Discrete-time control for switched positive systems with application to mitigating viral escape," *International Journal of Robust and Nonlinear Control*, vol. 21, no. 10, pp. 1093–1111, 2011.

[11] M. A. Nowak, R. M. May, and R. M. Anderson, "The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease," *AIDS*, vol. 4, no. 11, pp. 1095–1103, 1990.

[12] B. Adams, H. Banks, H. Kwon, and H. Tran, "Dynamic multidrug therapies for HIV: optimal and STI control approaches," *Mathematical Biosciences and Engineering*, vol. 1, pp. 223–241, 2004.

[13] A. S. Perelson, D. E. Kirschner, and R. de Boer, "Dynamics of HIV infection of CD4+ T cells," *Mathematical Biosciences*, vol. 114, no. 1, pp. 81–125, 1993.

[14] S. H. Bajaria, G. Webb, M. Cloyd, and D. Kirschner, "Dynamics of naive and memory CD4+ T lymphocytes in HIV-1 disease progression," *Journal of Acquired Immune Deficiency Syndromes*, vol. 30, no. 1, pp. 41–58, 2002.

[15] D. E. Kirschner, R. Mehr, and A. S. Perelson, "Role of the thymus in pediatric HIV-1 infection," *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, vol. 18, no. 2, pp. 95–109, 1998.

[16] P. Ye, D. E. Kirschner, and A. P. Kourtis, "The thymus during HIV disease: role in pathogenesis and in immune recovery," *Current HIV Research*, vol. 2, no. 2, pp. 177–183, 2004.

[17] D. Kirschner, G. F. Webb, and M. Cloyd, "Model of HIV-1 disease progression based on virus-induced lymph node homing and homing-induced apoptosis of CD4+ lymphocytes," *Journal of Acquired Immune Deficiency Syndromes*, vol. 24, no. 4, pp. 352–362, 2000.

[18] L. Rong and A. S. Perelson, "Modeling HIV persistence, the latent reservoir, and viral blips," *Journal of Theoretical Biology*, vol. 260, no. 2, pp. 308–331, 2009.

[19] Z. Shu, E. A. Hernandez-Vargas, and R. H. Middleton, "A math ematical study on immune activation and related dynamics in HIV infection," in *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC '11)*, vol. 2, pp. 3682–3687, 2011.

[20] I. B. Hogue, S. H. Bajaria, B. A. Fallert, S. Qin, T. A. Reinhart, and D. E. Kirschner, "The dual role of dendritic cells in the immune response to human immunodeficiency virus type 1 infection," *Journal of General Virology*, vol. 89, no. 9, pp. 2228–2239, 2008.

[21] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho, "HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time," *Science*, vol. 271, no. 5255, pp. 1582–1586, 1996.

[22] D. S. Callaway and A. S. Perelson, "HIV-1 infection and low steady state viral loads," *Bulletin of Mathematical Biology*, vol. 64, no. 1, pp. 29–64, 2002.

[23] A. Yates, J. Stark, N. Klein, R. Antia, and R. Callard, "Understanding the slow depletion of memory CD4+ T cells in HIV infection," *PLoS Medicine*, vol. 4, no. 5, Article ID e177, 2007.

[24] N. Dalal, D. Greenhalgh, and X. Mao, "A stochastic model for internal HIV dynamics," *Journal of Mathematical Analysis and Applications*, vol. 341, no. 2, pp. 1084–1101, 2008.

[25] N. I. Stilianakis, K. Dietz, and D. Schenzle, "Analysis of a model for the pathogenesis of AIDS," *Mathematical Biosciences*, vol. 145, no. 1, pp. 27–46, 1997.

[26] E. A. Hernandez-Vargas and R. H. Middleton, "Modeling the three stages in HIV infection," *Journal of Theoretical Biology*, vol. 320, pp. 33–40, 2013.

[27] J. M. Orenstein, "The macrophage in HIV infection," *Immunobiology*, vol. 204, no. 5, pp. 598–602, 2001.

[28] D. Kirschner and A. Perelson, "A model for the immune system response to HIV: AZT treatment studies," in *Mathematical Population Dynamics*, pp. 295–310, Wuerz, Winnepeg, Canada, 1995.

[29] M. Hadjiandreou, R. Conejeros, and V. S. Vassiliadis, "Towards a long-term model construction for the dynamic simulation of HIV infection," *Mathematical Biosciences and Engineering*, vol. 4, no. 3, pp. 489–504, 2007.

[30] C. J. Mode, *Applications of Monte Carlo Methods in Biology, Medicine and Other Fields of Science*, InTech, 2011.

[31] A. S. Fauci, G. Pantaleo, S. Stanley, and D. Weissman, "Immunopathogenic mechanisms of HIV infection," *Annals of Internal Medicine*, vol. 124, no. 7, pp. 654–663, 1996.

[32] S. Sungkanuparph, R. K. Groger, E. T. Overton, V. J. Fraser, and W. G. Powderly, "Persistent low-level viraemia and virological failure in HIV-1-infected patients treated with highly active antiretroviral therapy," *HIV Medicine*, vol. 7, no. 7, pp. 437–441, 2006.

[33] A. Sáez-Cirión, C. Bacchus, L. Hocqueloux et al., "Post-treatment HIV-1 controllers with a long-term Virological remission after the interruption of early initiated antiretroviral therapy ANRS VISCONTI study," *PLoS Pathogens*, vol. 9, Article ID e1003211, 2013.