Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 705821, 8 pages
doi:10.1155/2010/705821

Research Article

MHC Class II Binding Prediction—A Little Help from a Friend

Ivan Dimitrov,[1] Panayot Garnev,[1] Darren R. Flower,[2] and Irini Doytchinova[1]

[1] *Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st., 1000 Sofia, Bulgaria*
[2] *Life and Health Sciences, Aston University, Aston Triangle, Birmingham B4 7ET, UK*

Correspondence should be addressed to Irini Doytchinova, idoytchinova@pharmfac.net

Received 27 September 2009; Revised 20 January 2010; Accepted 22 February 2010

Academic Editor: Yongqun Oliver He

Copyright © 2010 Ivan Dimitrov et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Vaccines are the greatest single instrument of prophylaxis against infectious diseases, with immeasurable benefits to human wellbeing. The accurate and reliable prediction of peptide-MHC binding is fundamental to the robust identification of T-cell epitopes and thus the successful design of peptide- and protein-based vaccines. The prediction of MHC class II peptide binding has hitherto proved recalcitrant and refractory. Here we illustrate the utility of existing computational tools for in silico prediction of peptides binding to class II MHCs. Most of the methods, tested in the present study, detect more than the half of the true binders in the top 5% of all possible nonamers generated from one protein. This number increases in the top 10% and 15% and then does not change significantly. For the top 15% the identified binders approach 86%. In terms of lab work this means 85% less expenditure on materials, labour and time. We show that while existing caveats are well founded, nonetheless use of computational models of class II binding can still offer viable help to the work of the immunologist and vaccinologist.

1. Introduction

Vaccines continue to have an enormous and unprecedented positive impact on humanity and its wellbeing. Hundreds of millions of human lives have been saved since the first vaccine was discovered: Edward Jenner's smallpox vaccine in 1796 [1]. Yet the need to develop and deploy new vaccines has never been more urgent. Infectious disease causes about 25% of global deaths, particularly in children under five. The leading annual causes of death are 2.9 millions for tuberculosis; 2.5 million for diarrhoeal illnesses, especially rotaviruses; a rapidly escalating 2.3 million for HIV/AIDS; and 1.08 million deaths for malaria. There are no effective vaccines for HIV and Malaria, and the only vaccine available for tuberculosis is of limited utility. Consider also the 35 new, previously unknown, infectious diseases identified in the past 25 years: ebola, SARS, Dengue, West Nile fever, and potentially pandemic H5N1 influenza among them.

Historically, vaccines have been attenuated whole pathogen vaccines such as BCG for TB or Sabin's Polio vaccine. Issues of safety have led to the development of other strategies for vaccine development, separately focusing on antigen and epitope vaccines. The epitope is the minimal structure able to evoke an immune response. It is the immunological quantum that lies at the heart of immunity. Epitope-based vaccines have the advantage that many sequences able to induce autoimmunity or adverse reactions can be eliminated. Such vaccines are intrinsically safer: they contain no viable microorganisms and cannot induce microbial disease. However, several significant obstacles must be overcome before epitope-based vaccines can reach the market en masse. One such obstacle is MHC polymorphism.

Major histocompatibility complex (MHC) proteins, also known as human leukocyte antigens (HLA), are glycoproteins which bind within the cell short peptides, also called epitopes, derived from host and/or pathogen proteins, and present them at the cell surface for inspection by T-cells. T cell recognition is a fundamental mechanism of the adaptive immune system by which the host identifies and responds to foreign antigens [2].

There are two classes of MHC molecules: class I and class II. MHC class I molecules typically present peptides from proteins synthesized within the cell (endogenous processing pathway). MHC class II proteins primarily present peptides derived from endocytosed extracellular proteins (exogenous processing pathway). Both classes of MHC proteins are extremely polymorphic. More than 3500 molecules are listed in IMGT/HLA database [3]. MHC class I proteins are encoded by three loci: HLA-A, HLA-B, and HLA-C. MHC class II proteins also are encoded by three loci: HLA-DR, HLA-DQ, and HLA-DP. The peptide binding site of class I proteins has a closed cleft, formed by a single protein chain (α-chain) [4]. Usually, only short peptides of 8 to 11 amino acids bind in an extended conformation. In contrast, the cleft of class II proteins is open-ended, allowing much longer peptides to bind, although only 9 amino acids actually occupy the site. The class II cleft is formed by two separate protein chains: α and β [4]. Both clefts have binding pockets, corresponding to primary and secondary anchor positions on the binding peptide. The combination of two or more anchors is called a motif. The experimental determination of motifs for every allele is prohibitively expensive in terms of labor, time, and resources. The only practical and useable alternative is a bioinformatics approach.

Many bioinformatics methods exist to predict peptide-MHC binding [5]. Experimentally determined affinities data have formed the basis of many peptide-MHC binding prediction methods, able effectively to discriminate binding-from nonbinding peptides. Such methods include so-called *motifs*, as well as highly sophisticated computer science algorithms—artificial neural networks [6], HMMs [7], and support vector machines (SVMs) [8]—and methods derived from computational chemistry, such as QSAR analysis [9] and structure-based approaches [10].

MHC-binding motifs are an easily understood method of epitope identification. They generate many false-positives and many false-negatives. SVMs are based on statistical theory that seeks to induce a dichotomy of distinct classes. HMMs model systems by assuming them to be Markov processes with unknown parameters. An HMM profile can determine those sequences which exhibit binder-like characteristics. QSAR techniques can refine the peptide interactions within the MHC class I groove by optimizing individual residue-to-residue pairs.

Several sophisticated methods have been created to resolve the dynamic variable-length problem inherent within the class II prediction. Such methods include an iterative "meta-search" algorithm, Ant Colony search, Gibbs sampling algorithm, and multiobjective evolutionary algorithm. Certain new approaches have significantly outperformed more traditional methods [11]. No single method yet proposes a wholly satisfying and satisfactory solution to this dilemma. The efficiency demonstrated by these algorithms is often very different for different class II alleles and there is little overlap between peptide rankings generated by these methods. This has resulted in much pessimism regarding the usefulness of a computational approach. Here, we seek to address this issue.

In the present study, a set of 167 proteins containing 4540 epitopes binding to HLA-DRB1 alleles was compiled and used to test the predictive ability of several publically-available servers for MHC binding predictions. Our aim was not to compare the performance of the available servers, per se. This has been undertaken several times already. Rather, in the context of tasks regularly undertaken by immunologists and vaccinologists, we wish to illustrate the utility of existing computational tools for in silico prediction of peptides binding to class II MHCs. We show that while existing caveats are wellfounded, nonetheless use of computational models of class II binding can still yield viable help to the work of immunologists and vaccinologists.

2. Materials and Methods

2.1. Test Set Used in the Study. At the time of evaluation (December 2009), the Immune Epitope Database [12] contained 10 925 peptides of different length binding to HLA-DRB1 alleles. For the purpose of our study we extracted only binders annotated with an identified source protein. After removing the duplicate sequences and proteins containing unknown amino acids, the final set consisted of 4540 binders, belonging to 167 proteins, and binding to 12 widely spread HLA-DRB1 alleles. The alleles used in the study were DRB1*0101 (2051 binders), DRB1*0301 (190 binders), DRB1*0401 (392 binders), DRB1 *0404 (159 binders), DRB1*0405 (244 binders), DRB1*0701 (336 binders), DRB1*0802 (153 binders), DRB1*0901 (160 binders), DRB1*1101 (275 binders), DRB1*1201 (24 binders), DRB1*1302 (243 binders), and DRB1*1501 (313 binders).

2.2. Servers for MHC Class II Binders Prediction Used in the Study. The servers for MHC class II binding prediction used in our assessment were selected on the basis of matching to the following criteria: computational or machine-learning method-based, free web access, and the ability to predict binding to at least 10 of the 12 HLA-DRB1 alleles considered in the study (Table 1).

ProPred [13] predicts MHC class II binding peptides using the quantitative matrix-based pocket profiles of Sturniolo et al. [14]. RANKPEP [15] uses position-specific scoring matrices (PSSM) or profiles which represent the observed sequence-weighted frequency of all amino acids in every position of a sequence alignment. IEDB-ARB [16] is a matrix-based prediction method where the peptide binding score is calculated by multiplying the relative contribution coefficients for each amino acid at each peptide position. The IEDB-SMM align method [17] is based on an integrated alignment and motif identification algorithm and predicts directly peptide binding affinities. MHC2Pred [18] is a SVM-based prediction server. EpiTOP [19] is a newly developed method for MHC class II binding prediction based on proteochemometrics [20]. It is a matrix-based method which considers both peptide and protein binding site amino acids contributions. NetMHCII and NetMHCIIpan are ANN-based methods. NetMHCIIpan takes into account both peptide and MHC sequence information [21].

Most of the servers do not predict binding to all DRB1 alleles used in the test sets. Only servers IEDB, NetMHCIIpan, and EpiTOP make predictions for all 12 DRB1 alleles. Obviously, each server was only evaluated using the alleles it predicts.

Table 1: Servers for MHC class II binders prediction used in the study.

Server	Method	URL
NetMHCII	ANN[a]	http://www.cbs.dtu.dk/services/NetMHCII/
NetMHCIIpan	ANN	http://www.cbs.dtu.dk/services/NetMHCIIpan/
ProPred	QM[b]	http://www.imtech.res.in/raghava/propred/
RANKPEP	QM	http://bio.dfci.harvard.edu/RANKPEP/
IEDB-ARB	QM	http://tools.immuneepitope.org/analyze/html/mhc_II_binding.html
IEDB-SMM	QM	http://tools.immuneepitope.org/analyze/html/mhc_II_binding.html
EpiTOP	QM	http://www.pharmfac.net/EpiTOP/
MHC2Pred	SVM[c]	http://www.imtech.res.in/raghava/mhc2pred/

[a]ANN: artificial neural networks, [b]QM: quantitative matrix, [c]SVM: support vector machine.

3. Performance Evaluation

The evaluation was performed under conditions similar to those an experimental immunologist might use: the complete protein sequences were submitted to a server and the results recorded. Five cutoffs were used to categorise the result: top 5%, 10%, 15%, 20%, and 25% of the predicted binding nonamers. ProPred only returns the top 10% of predictions; for RANKPEP this limitation is 20%. An identified binding peptide was considered to be any nonamer identical in sequence to a nonamer from the set of known binders, and originating from the same protein. Identified binders are shown as a percentage of all binders (*sensitivity*). Although many of the methods give quantitative predictions, in the evaluation study they were used as classification methods.

4. Results

*4.1. HLA-DRB1*0101 Binders Predictions.* The test subset of peptides binding to HLA-DRB1*0101 consisted of 2051 binders. Four of the servers (NetMHCII, NetMHCpan, RANKPEP, and EpiTOP) recognize more than 60% of them in the top 10% (Figure 1(a)). The number of the identified binders increases in the next cutoff steps reaching 93% by NetMHCpan, 91% by EpiTOP, and 88% by NetMHCII and RANKPEP at the top 25%.

*4.2. HLA-DRBI*0301 Binders Predictions.* One hundred and ninety binders from the test set bind to HLA-DRB1*0301. More than half of them are recognized by NetMHCpan, NetMHCII, and ProPred even at the top 5% of the predicted best binders. Sensitivity above 80% is achieved by NetMHCpan, NetMHCII, and EpiTOP at cutoff of 15% (Figure 1(b)).

*4.3. HLA-DRBI*0401 Binders Predictions.* The subset of HLA-DRB1*0401 binders consisted of 392 binders. Most of the servers present well at the top 5% level recognizing more than a half of the binders (Figure 1(c)). At the 20% level NetMHCII, and NetMHCIIpan present best identifying 96% and 95% of the binders, respectively, followed by EpiTOP and RANKPEP (91%).

*4.4. HLA-DRBI*0404 Binders Predications.* The binders to HLA-DRB1*0404 in the test set were 159. Only NetMHCIIpan achieves more than 60% sensitivity at the top 5% cutoff. More than 80% of the known binders are recognized by NetMHCIIpan at the top 10% level and by RANKPEP, EpiTOP, and NetMHCII at the top 15% (Figure 1(d)).

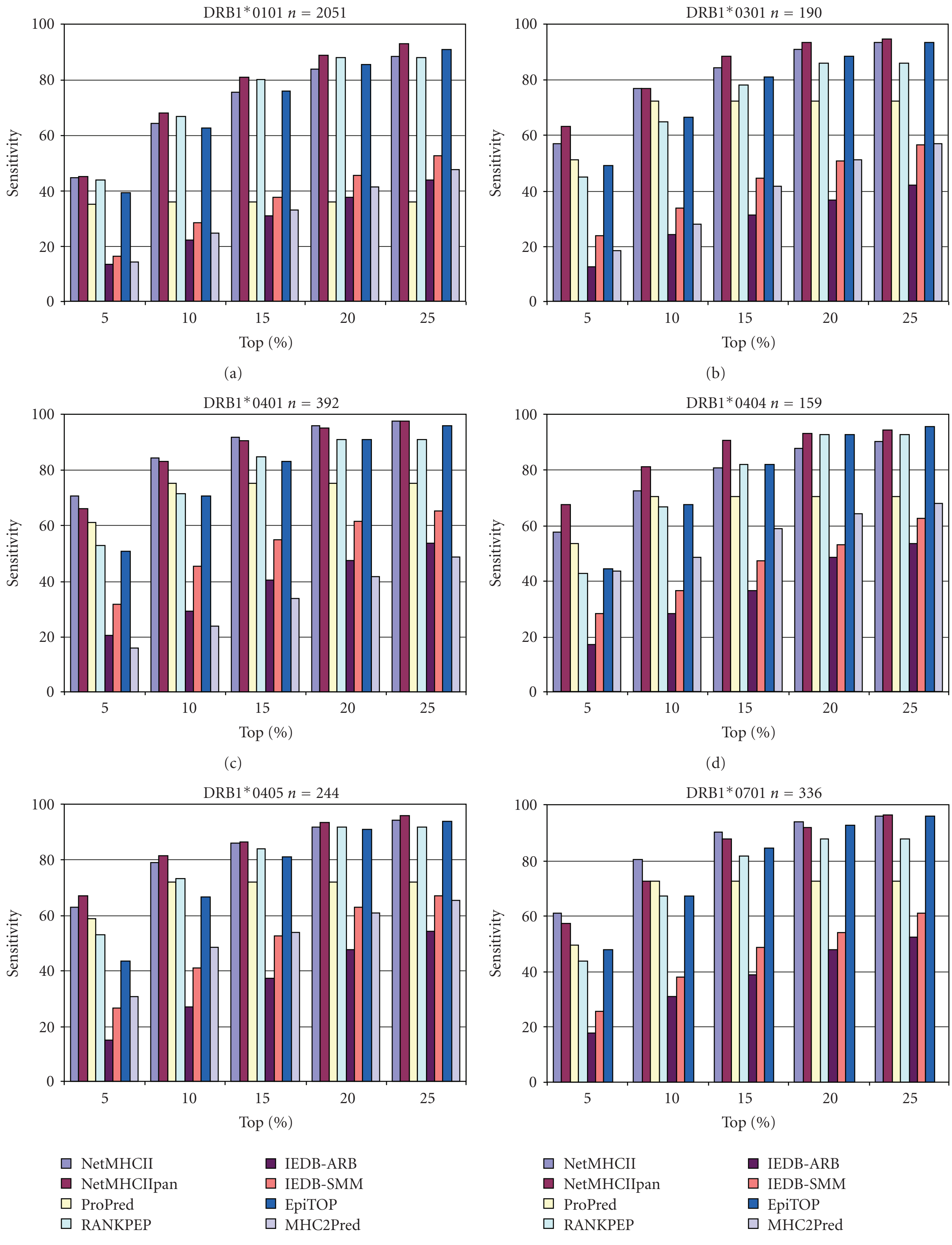
*4.5. HLA-DRB1*0405 Binders Predictions.* The test set contained 244 binders to HLA-DRB1*0405. The performance of the servers on this allele was very similar to HLA-DRB1*0404. NetMHCIIpan and NetMHCII recognize more than 60% of the binders in the top 5% of the predicted best binders. Sensitivity of 80% is achieved by NetMHCIIpan at the top 10% level and by RANKPEP, EpiTOP, and NetMHCII at the top 15% (Figure 1(e)).

*4.6. HLA-DRB1*0701 Binders Predictions.* Three hundred forty four are the binders to HLA-DRB1*0701 in the test set. NetMHCII performs best here identifying more than 60% of the known binders at the top 5% level and 80% of them at the top 10% (Figure 1(f)). Sensitivity of 80% is achieved by NetMHCIIpan, RANKPEP, and EpiTOP at the top 15%. MHC2Pred does not predict affinity to this allele.

*4.7. HLA-DRB1*0802 Binders Predictions.* The binders to HLA-DRB1*0802 in the test set were 153. NetMHCIIpan, NetMHCII, and ProPred achieve more than 60% sensitivity at the top 5% cutoff (Figure 1(g)). A sensitivity higher than 80% has NetMHCIIpan at top 10% level, and NetMHCII and EpiTOP—at top 15%. RANKPEP is not trained to predict binders to this allele

*4.8. HLA-DRB1*0901 Binders Predictions.* The test set contained 160 binders to HLA-DRB1*0901. NetMHCIIpan recognizes 68% of the known binders to this allele in the top 5% (Figure 1(h)). Sensitivity of 80% is achieved by NetMHCII, NetMHCIIpan, EpiTOP, and RANKPEP at the top 15% level. ProPred does not make predictions for this allele.

*4.9. HLA-DRB1*1101 Binders Predictions.* Two hundred seventy five peptides in the test set happen to bind to HLA-DRB1*1101. NetMHCII and NetMHCIIpan recognize 60%



(a)

(b)

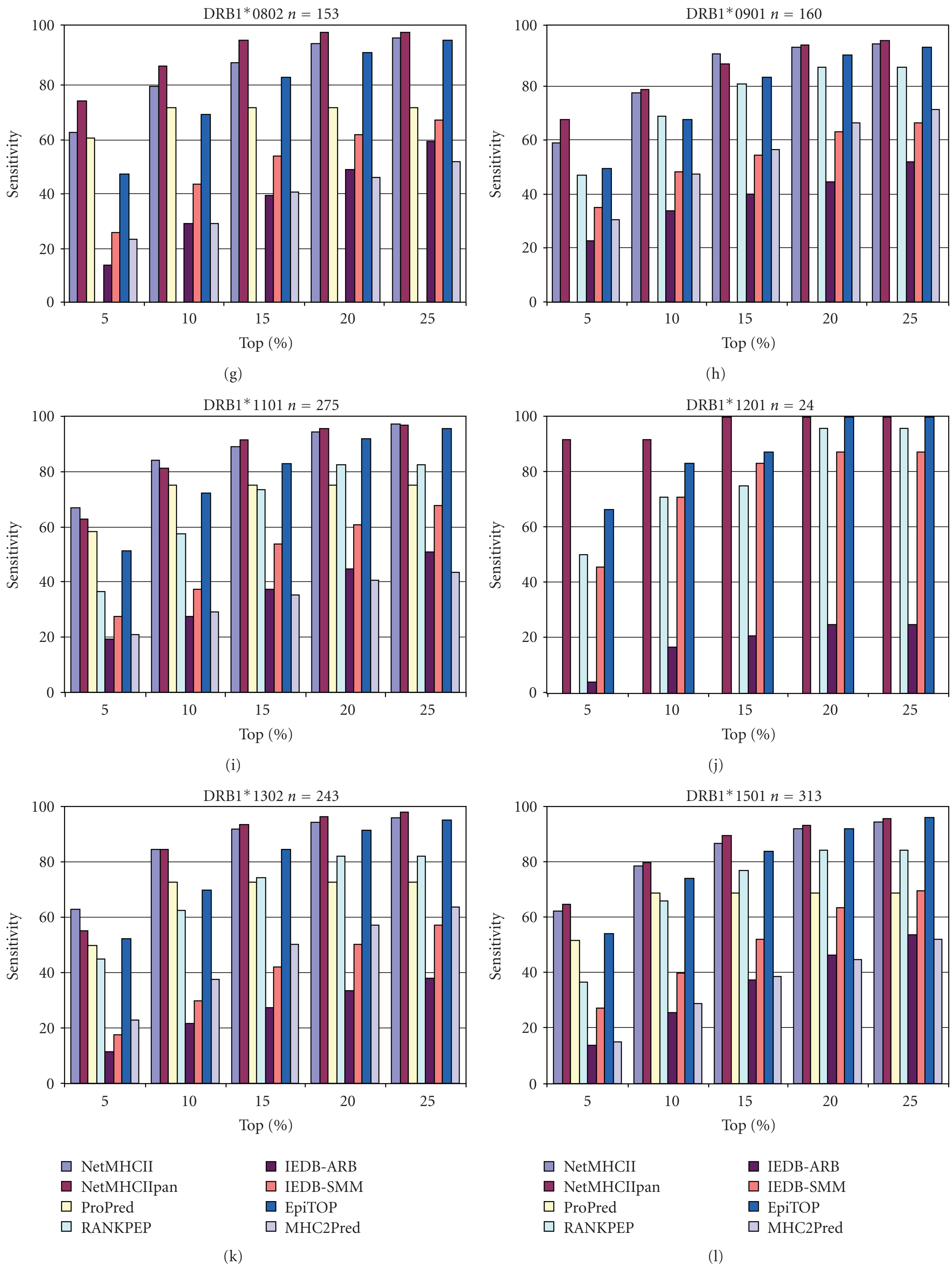
(c)

(d)

(e)

(f)

Figure 1: Continued.



(g) (h) (i) (j) (k) (l)

Figure 1: Number of identified binders (*sensitivity*) in the top 5%, 10%, 15%, 20% and 25% of all overlapping nonamers generated from a protein: (a) DRB1*0101, (b) DRB1*0301, (c) DRB1*0401, (d) DRB1*0404, (e) DRB1*0405, (f) DRB1*0701, (g) DRB1*0802, (h) DRB1*0901, (i) DRB1*1101, (j) DRB1*1201, (k) DRB1*1302, (l) DRB1*1501.

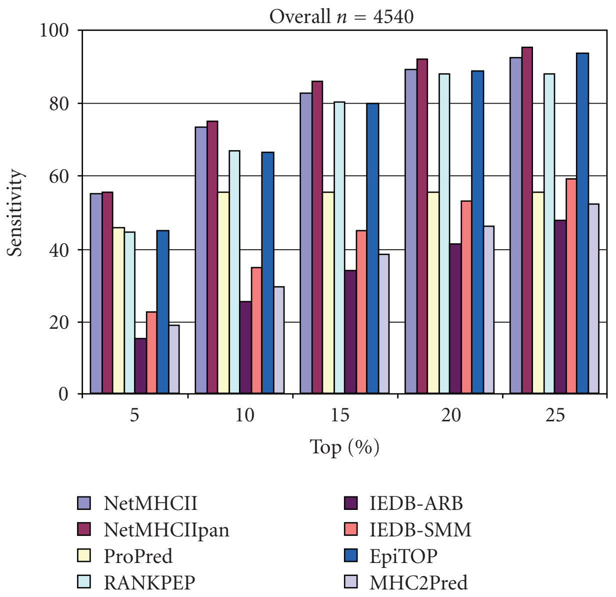


Figure 2: Overall HLA-DRB1 binders prediction.

of them at the top 5% and 80% of them at the top 10% of the predicted best binders (Figure 1(i)).

*4.10. HLA-DRB1*1201 Binders Predictions.* Only 24 peptides are the binders to HLA-DRB1*1201 in the test set. NetMHCII performs best here identifying 92% of the known binders at the top 5% level and 100% of them at the top 15% (Figure 1(j)). Second best is EpiTOP with sensitivity of 83% at the top 10%. NetMHCII, ProPred, and MHC2Pred do not predict affinity to this allele.

*4.11. HLA-DRB1*1302 Binders Predictions.* The binders to HLA-DRB1*1302 in the test set were 243. Only NetMHCIIpan achieves more than 60% sensitivity at the top 5% cutoff (Figure 1(k)). A sensitivity higher than 80% has NetMHC and NetMHCIIpan at top 10% level, and EpiTOP—at top 15%.

*4.12. HLA-DRB1*1501 Binders Predictions.* The subset of peptides binding to HLA-DRB1*1501 consisted of 313 binders. Sixty percent of them are recognized by NetMHCII and NetMHCIIpan at the top 5% cutoff (Figure 1(l)). NetMHCIIpan reaches 80% sensitivity at the top 10%, while NetMHCII and EpiTOP—at the top 15%.

4.13. Overall HLA-DRB1 Binders Predictions. Half of the binders are identified within the top 5% by NetMHCIIpan and NetMHCII (Figure 2). Sensitivity of 80% is achieved by NetMHCIIpan, NetMHCII, RANKPEP, and EpiTOP within the top 15%. This performance is maintained in the top 20% and top 25% and top 25%; see supplementary material available online at doi: 10.1165/2010/705821.

5. Discussion

The peptides presented by MHC class II molecules are derived predominantly from extracellular proteins (not cytosolic as in MHC class I), which are mainly of bacterial origin. They are endocytosed by professional antigen presenting cells (APCs) such as dendritic cells, macrophages and B-cells, digested in lysosomes by cathepsin S [22], and bound by class II molecules in subcellular vesicles. Then the complex peptide-class II molecule is expressed on the cell surface and interacts exclusively with CD4$^+$T cells (helper T-cells, T$_H$C). T$_H$ cells help to trigger an appropriate immune response which may include localized inflammation and swelling due to recruitment of phagocytes or may lead to a full-force antibody-mediated immune response due to the activation of B cells.

MHC binding is the stage of antigen presentation which we understand best in both the class I and class II pathways. It is also the most discriminating stage within the presentation-recognition pathways. That is why most of the methods for T-cell epitope prediction are in practice methods for MHC binding prediction. The successful prediction of MHC class II binding is more difficult than the successful prediction of class I binding. The main difficulty is the unrestricted length of class II epitopes. Compared with MHC class I binders, which are limited up to 11 amino acids, though sometimes longer, the open-ended class II binding site does not constrain peptide lengths, allowing binding of peptides consisted of up to or more than 25 amino acids. However, as X-ray structures show, the class II binding site is always occupied by nine amino acids, with the rest of the peptide protruding at both sides. Thus, class II prediction methods need to identify the binding nonamer for each sequence and then develop a predictive model. Most of the models consider the principle "one nonamer per binding sequence". Experimental data, however, suggests the possible existence of multiple registers with different nonameric core regions within a binding peptide, each serving as recognition sites for MHC class II molecules [23]. Our previous work indicated that when an easily distinguished good binder is not available in the peptide sequence, the binding affinity is a degenerate average of affinities from several binding subsequences [24]. The multiple registers of binding peptides and the degenerate recognition might explain the lower predictive ability for MHC class II compared to those for MHC class I [25].

The present comparative study was provoked by two emergent themes in the immunoinformatics literature concerning the development of T-cell epitope predictive methods. The first includes papers describing new predictive methods. Most such studies tend to favour the method they invent and this inevitably influences the choice of test data and the way in which that test is conducted. Thus, the test set is never truly independent and the reason for this is the immanent bias associated with the process of selecting test data.

The second trend, which inspired us to perform this study, was the prevalent negative attitude of many scientists to the efforts of the immunoinformatic community, which has sought to develop methods that will facilitate experimental lab work reducing significantly the practitioners' laborious times and resource. Where data are sufficiently abundant, methods aimed at predicting class I MHC binding seem to work well [26]. However, other types of prediction have the reputation of working poorly. for example, the structure-driven prediction of class I and class II T-cell epitopes [27]. More recently, several comparative studies have shown that, in particular, the prediction of class II T-cell epitopes is suboptimal [28–30]. This has led to unnecessary pessimism regarding the utility of such approaches; here we seek to redress this unnecessarily cautious attitude and perception.

For different reasons, reliable prediction across the board remains elusive, and will continue so for some time. Nonetheless, even extant predictive informatic methodologies, when applied shrewdly and combined synergistically with other approaches, can deliver clear and useful benefits.

Methods are limited by data. What is needed are properly designed data sets which can properly sample and explore the multidimensional space accessed by all congeneric molecules under examination. No data-driven method can go beyond the training data: all methods are better at interpolating than extrapolating. It is only by having excellent and general data that we can hope for general and excellent models.

Except in rare cases, data is usually multi-dimensional, and each dimension will typically be correlated, to a greater or lesser extent, with one or more other dimensions. Together, these many dimensions delineate a space: a space of structural variation or variation of properties. If our data is itself of sufficient quality and provides a good enough coverage of the space, then straightforward methods drawn from, say, computer science—of which there are indeed very many—are now of sufficient accuracy to generate models of high predictive accuracy.

The quality, quantity, and availability of data must increase and improve, particularly for class II. Nonetheless, existing methods built on extant data can, in spite of its inherent imperfections, still be useable and useful, as we show here. Prediction is captive to its underlying data. Bias within the data places strict limitations upon the interpretability and generality of models derived from it. In general, for MHC-peptide binding experiments, the sequences of peptides studied are very biased in terms of amino acid composition, often favouring hydrophobic sequences. This arises, in part, from preselection processes that result in self-reinforcement. Binding motifs are often used to reduce the experimental burden of epitope discovery. Very sparse sequence patterns are matched and the corresponding subset of peptides tested, with an enormous reduction in sequence diversity.

In this study, we choose a test set of 4540 known binders to HLA-DRB1 alleles from the Immune Epitope Database [12]. The binders were of different length and originate from known source proteins. The number of the source proteins was 167. Peptides bound to 12 widely spread HLA-DRB1 alleles. The evaluation was performed under conditions similar to those which an experimental immunologist might use: the complete sequence of a protein of interest is submitted to an available web server and the results recorded. Five thresholds were used: top 5%, 10%, 15%, 20%, and 25% of all overlapping nonamers generated from a protein. The identified binders were presented as a percent of each allele binders (*sensitivity*) used in the study.

The results from our evaluation are unexpectedly encouraging. Half of the known binders are identified within the top 5% by two servers: NetMHCIIpan and NetMHCII (Figure 2). For some alleles, the sensitivity at this level reaches, 92% (NetMHCIIpan for HLA-DRB1*1201). The sensitivity increases in the top 10% and 15% and then does not alter significantly. Most of the servers achieve sensitivity of 80% at the top 15%. In terms of lab work, this means 85% less expenditure on expended on materials, labour, and time. Apart from NetMHCIIpan and NetMHCII, two other servers—RANKPEP and EpiTOP—perform well. The moderate performance of IEDB-ARB, IEDB-SMM, and MHC2Pred, which are well known and widely used servers for MHC class II binding prediction, is surprising. One possible explanation could be the high levels of specificity, predefined in this evaluation through the cutoffs of top 5% to top 25% of the predicted best binders. The aim of these high levels of specificity is to avoid the great number of false positives often generated by quantitative predictions.

The results we present here for MHC binding prediction illustrate the usefulness of computational tools for the everyday work of the immunologist. These results go a long way to refuting - and refuting eloquently - the apparent skepticism among many experimental immunologists and vaccinologists regarding efforts by immunoinformatics and immunoinformaticians to design and develop highly predictive computational tools for the in silico identification of T-cell epitopes. Accurate prediction remains vital for the future of vaccine informatics and for vaccinology as a whole. It is important to realize what can be and what cannot be done.

In future work, we will use the specific results and general know-how generated in this study to inform data-fusion approaches to improve class II peptide-MHC binding prediction. In particular, we will explore the use of optimised voting algorithms to generate a viable meta-predictor, which unites the output of several prediction methods in an intelligent manner so that the combined output is more accurate and more reliable than any individual prediction program. Such approaches have been widely employed in other areas and even in immunoinformatics: Trost et al. have addressed class I binding [31], while Karpenko et al. have used this approach to predict class II MHCs binding [32]. We will seek to capitalise upon the as-yet-unrealised potential of such approaches.

What vaccine informatics offers are tools that can become important components of a deeper, broader experimental and clinical endeavour. The models we explore above are tools of true utility replete with practical real-world applications. Vaccinology and immunology, as disciplines, need only embrace such methods; in their turn such techniques will liberate the vaccinologist and immunologist from the drudgery of uninformed experimentation, allowing them to design better, faster, smarter ways of discovery of new reagents, diagnostics, and vaccines.

Acknowledgment

The research was supported by The National Science Fund of Ministry of Education and Science, Bulgaria (Grant 02-115/2008). The supplementary material provides detailed results for HLA-DRB1 binding prediction using a test set of 4540 known binders.

References

[1] D. R. Flower, "Vaccines: their place in history," in *Bioinformatics for Vaccinology*, pp. 1–54, Wiley-Blackwell, Oxford, UK, 2008.

[2] D. R. Flower, "Vaccines: how they work," in *Bioinformatics for Vaccinology*, pp. 73–112, Wiley-Blackwell, Oxford, UK, 2008.

[3] J. Robinson, M. J. Waller, P. Parham, et al., "IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex," *Nucleic Acids Research*, vol. 31, no. 1, pp. 311–314, 2003.

[4] C. A. Janeway, P. Travers, M. Walport, and J. D. Capra, "The recognition of antigen," in *Immunobiology: The Immune System in Health and Disease*, pp. 79–194, 1999.

[5] D. R. Flower, "Vaccines: data driven prediction of binders, epitopes and immunogenicity," in *Bioinformatics for Vaccinology*, pp. 167–216, Wiley-Blackwell, Oxford, UK, 2008.

[6] K. Gulukota and C. DeLisi, "Neural network method for predicting peptides that bind major histocompatibility complex molecules," *Methods in Molecular Biology*, vol. 156, pp. 201–209, 2001.

[7] H. Noguchi, R. Kato, T. Hanai, et al., "Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules," *Journal of Bioscience and Bioengineering*, vol. 94, no. 3, pp. 264–270, 2002.

[8] J. Wan, W. Liu, Q. Xu, Y. Ren, D. R. Flower, and T. Li, "SVRMHC prediction server for MHC-binding peptides," *BMC Bioinformatics*, vol. 7, article 463, 2006.

[9] I. A. Doytchinova, V. Walshe, P. Borrow, and D. R. Flower, "Towards the chemometric dissection of peptide—HLA-A*0201 binding affinity: comparison of local and global QSAR models," *Journal of Computer-Aided Molecular Design*, vol. 19, no. 3, pp. 203–212, 2005.

[10] M. N. Davies, C. K. Hattotuwagama, D. S. Moss, M. G. B. Drew, and D. R. Flower, "Statistical deconvolution of enthalpic energetic contributions to MHC-peptide binding affinity," *BMC Structural Biology*, vol. 6, article 5, 2006.

[11] J. Salomon and D. R. Flower, "Predicting Class II MHC-Peptide binding: a kernel based approach using similarity scores," *BMC Bioinformatics*, vol. 7, article 501, 2006.

[12] C. P. Toseland, D. J. Taylor, H. McSparron, et al., "AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical and cellular data," *Immunome Research*, vol. 1, article 4, 2005.

[13] H. Singh and G. P. S. Raghava, "ProPred: prediction of HLA-DR binding sites," *Bioinformatics*, vol. 17, no. 12, pp. 1236–1237, 2001.

[14] T. Sturniolo, E. Bono, J. Ding, et al., "Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices," *Nature Biotechnology*, vol. 17, no. 6, pp. 555–561, 1999.

[15] P. A. Reche, J.-P. Glutting, H. Zhang, and E. L. Reinherz, "Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles," *Immunogenetics*, vol. 56, no. 6, pp. 405–419, 2004.

[16] H.-H. Bui, J. Sidney, B. Peters, et al., "Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications," *Immunogenetics*, vol. 57, no. 5, pp. 304–314, 2005.

[17] M. Nielsen, C. Lundegaard, and O. Lund, "Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method," *BMC Bioinformatics*, vol. 8, article 238, 2007.

[18] http://www.imtech.res.in/raghava/mhc2pred.

[19] http://www.pharmfac.net/EpiTOP/.

[20] I. Dimitrov, P. Garnev, D. R. Flower, and I. Doytchinova, "Peptide binding to the HLA-DRB1 supertype: a proteochemometrics analysis," *European Journal of Medicinal Chemistry*, vol. 45, no. 1, pp. 236–243, 2010.

[21] M. Nielsen, C. Lundegaard, T. Blicher, et al., "Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan," *PLoS Computational Biology*, vol. 4, no. 7, Article ID e1000107, 2008.

[22] L. Delamarre, M. Pack, H. Chang, I. Mellman, and E. S. Trombetta, "Differential lysosomal proteolysis in antigen-presenting cells determines antigen fate," *Science*, vol. 307, no. 5715, pp. 1630–1634, 2005.

[23] J. C. Tong, G. L. Zhang, T. W. Tan, J. T. August, V. Brusic, and S. Ranganathan, "Prediction of HLA-DQ3.2β ligands: evidence of multiple registers in class II binding peptides," *Bioinformatics*, vol. 22, no. 10, pp. 1232–1238, 2006.

[24] I. A. Doytchinova and D. R. Flower, "Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction," *Bioinformatics*, vol. 19, no. 17, pp. 2263–2270, 2003.

[25] V. Brusic, G. Rudy, M. Honeyman, J. Hammer, and L. Harrison, "Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network," *Bioinformatics*, vol. 14, no. 2, pp. 121–130, 1998.

[26] B. Peters, H. H. Bui, S. Frankild, et al., "A community resource benchmarking predictions of peptide binding to MHC-I molecules," *PLoS Computational Biology*, vol. 2, no. 6, article e65, 2006.

[27] B. Knapp, U. Omasits, S. Frantal, and W. Schreiner, "A critical cross-validation of high throughput structural binding prediction methods for pMHC," *Journal of Computer-Aided Molecular Design*, vol. 23, no. 5, pp. 301–307, 2009.

[28] Y. EL-Manzalawy, D. Dobbs, and V. Honavar, "On evaluating MHC-II binding peptide prediction methods," *PLoS ONE*, vol. 3, no. 9, article e3268, 2008.

[29] 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, supplement 12, article S22, 2008.

[30] U. Gowthaman and J. N. Agrewala, "In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion," *Journal of Proteome Research*, vol. 7, no. 1, pp. 154–163, 2008.

[31] B. Trost, M. Bickis, and A. Kusalik, "Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools," *Immunome Research*, vol. 3, no. 1, article 5, 2007.

[32] O. Karpenko, L. Huang, and Y. Dai, "A probabilistic meta-predictor for the MHC class II binding peptides," *Immunogenetics*, vol. 60, no. 1, pp. 25–36, 2008.