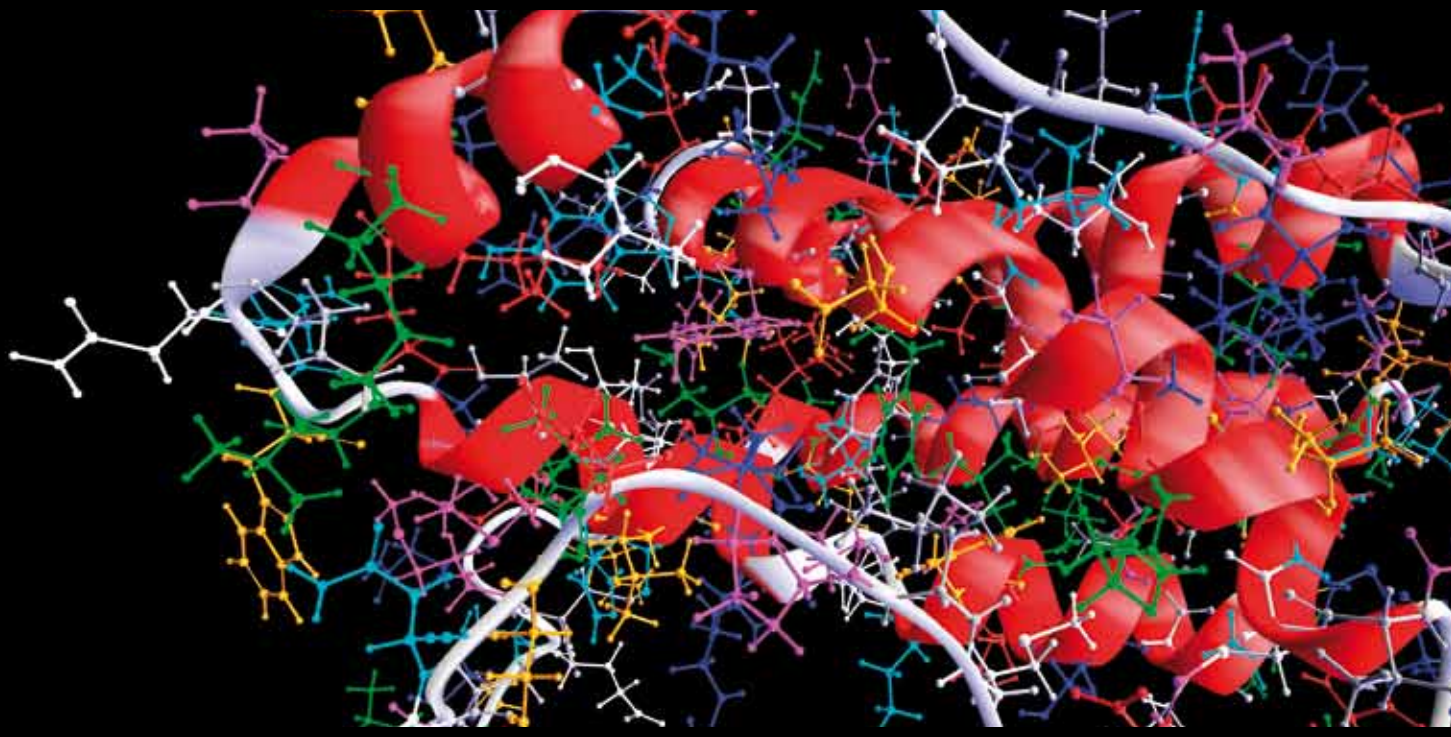
Computational and Mathematical Methods in Medicine

Phage Display Informatics

Guest Editors: Jian Huang, Ratmir Derda, and Yanxin Huang



Phage Display Informatics

Computational and Mathematical Methods in Medicine

Phage Display Informatics

Guest Editors: Jian Huang, Ratmir Derda, and Yanxin Huang

Editorial Board

Contents

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 698395, 2 pages
http://dx.doi.org/10.1155/2013/698395

Editorial

Phage Display Informatics

Jian Huang,[1] Ratmir Derda,[2] and Yanxin Huang[3]

[1] *Center of Bioinformatics (COBI), Key Laboratory for NeuroInformation of Ministry of Education, University of Electronic Science and Technology of China, Chengdu 610054, China*
[2] *Department of Chemistry and Alberta Glycomics Centre, University of Alberta, Edmonton, AB, Canada T6G 2G2*
[3] *National Engineering Laboratory for Druggable Gene and Protein Screening, Northeast Normal University, Changchun 130024, China*

Correspondence should be addressed to Jian Huang; hj@uestc.edu.cn

Received 22 August 2013; Accepted 22 August 2013

Phage display is an efficient laboratory technique that can be used to screen for specific peptides and proteins displayed on the surface of bacteriophage. Since Professor George Smith of the University of Missouri pioneered the powerful and flexible method in 1980s [1], it has been adapted and improved by many scientists from various fields. For example, the sequence displayed on the coat proteins of phage has been extended from random peptides to protein fragments, enzymes, antibodies, and even the whole peptidome of a given species [2]; the way of panning has been expanded from *in vitro* to *in vivo* [3]; the platform for screening has been extended from plates and beads to microfluidic devices [4]. In addition to the development of "hardwares" of phage display, researchers in closely relevant fields have also witnessed the birth and burst of "softwares" for managing enormous amounts of data on phage display and for making biological discoveries or predictions [5, 6]. With the spread of phage display technique and the progress of its "hardwares" and "softwares," it has made a great impact on modern medicine. For instance, phage display has been widely used for epitope mapping, analysis of protein-protein interactions, prediction of drug target, and identification of enzyme substrates and inhibitors. Some antibodies and peptides derived from phage display technology have been developed into new drugs approved by FDA; others have shown promise for the development of diagnostics, vaccines, and the targeted delivery of therapeutics. In these achievements, informatics means play an increasingly important role.

In this special issue, we take an interest in the investigation of computational and mathematical methods and their applications in all fields using phage display.

For both experimental biologists and computational biologists, mapping conformational B-cell epitopes is a very challenging task. The paper "*Bioinformatics resources and tools for conformational B-cell epitope prediction*" contributed by P. Sun et al. summarized the recent advance of bioinformatics resources and tools for the prediction of conformational B-cell epitopes. According to their review, the prediction methods based on the experimental results of phage display have become one major category of all algorithms. B. He et al. panned the Ph.D.-12 phage display peptide library against metuximab, a new drug for radioimmunotherapy of hepatocellular carcinoma approved by the State Food and Drug Administration of China in 2005, in the paper "*Epitope mapping of metuximab on CD147 using phage display and molecular docking.*" After cleaning their phage display data computationally, they predicted for the first time the complete epitope recognized by metuximab based on the analyses of mimotopes. Very interestingly, the prediction based on phage display largely overlapped with their docking result and the CD147-CD147 interfaces in the CD147 crystal structure. Consequently, they proposed that blocking the formation of CD147 dimer might be an important mechanism of metuximab function. The study by B. He et al. demonstrates that the prediction of conformational B-cell epitopes based on phage display is a cheap and quick strategy with an acceptable accuracy.

Though phage display was born for biomedicine studies, it has already gone beyond this field. For example, it has shown its power in the research for new material, new energy, environmental protection, and agriculture. R. Kushwaha et al. reviewed discoveries via phage display that impacted the use

of agricultural products in *"Uses of phage display in agriculture: a review of food-related protein-protein interactions discovered by biopanning over diverse baits."* Some parts of this review are relevant to medicine and new energy. For instance, the application of phage display in the studies of food allergy and biofuel production was highlighted. Moreover, the utilization of phage display in the defense of plants against herbivores and microbes was discussed. It was expected that phage display and relevant computational methods would become more popular in the agricultural research. Indeed, in another paper "*Uses of phage display in agriculture: sequence analysis and comparative modeling of late embryogenesis abundant client proteins suggest protein-nucleic acid binding functionality.*" by R. Kushwaha et al., sequence analysis and homology modeling were used to study 21 client proteins identified by phage display. The results from this initial computational study would guide their future efforts to uncover the protein protective mechanisms of plant seeds during heat stress.

As we mentioned previously, the blueprint of phage display proposed by Professor George Smith has inspired many scientists to adapt and improve this technique. Different phages and various coat proteins have been tested to construct new phage display systems. As the genomes of hundreds of phages have been sequenced, identification of their virion proteins will be helpful for the development of new phage display systems. P.-M. Feng et al. presented a Naïve Bayes-based method that can predict phage virion proteins using amino acid composition and dipeptide composition in *"Naïve bayes classifier with feature selection to identify phage virion proteins."* In their jackknife test, the classifier achieved an accuracy of 79.15% to divide phage virion and non-virion proteins, which were superior to other state-of-the-art methods.

Using next-generation sequencing techniques to enable cost-effective high-throughput analysis is a new trend in phage display technology. However, the trend suffers from errors in deep sequencing data, which may exceed 1%. W. Matochko et al. proposed a linear algebra framework for analyzing errors in a 7-mer peptide library with a medium scale sequenced by Illumina method in *"Error analysis of deep sequencing of phage libraries: peptides censored in sequencing."* As technical capabilities and depth of sequencing increases, the method would be applicable to larger libraries as well.

In summary, the six papers in this volume involve in various aspects of informatics tools and their applications in several fields using phage display technique. As a snapshot of phage display in the information age, it demonstrates that phage display in the 21st century is being transformed from a purely lab-based science to an information science as well, which can make it even powerful. With the rapid development of "hardwares" and "softwares" of phage display and information technology, we can even expect an in silico phage display system in future.

Jian Huang
Ratmir Derda
Yanxin Huang

References

[1] G. P. Smith, "Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface," *Science*, vol. 228, no. 4705, pp. 1315–1317, 1985.

[2] H. B. Larman, Z. Zhao, U. Laserson et al., "Autoantigen discovery with a synthetic human peptidome," *Nature Biotechnology*, vol. 29, no. 6, pp. 535–541, 2011.

[3] R. Pasqualini and E. Ruoslahti, "Organ targeting in vivo using phage display peptide libraries," *Nature*, vol. 380, no. 6572, pp. 364–366, 1996.

[4] K. Cung, R. L. Slater, Y. Cui et al., "Rapid, multiplexed microfluidic phage display," *Lab on a Chip*, vol. 12, no. 3, pp. 562–565, 2012.

[5] J. Huang, B. Ru, and P. Dai, "Bioinformatics resources and tools for phage display," *Molecules*, vol. 16, no. 1, pp. 694–709, 2011.

[6] J. Huang, B. Ru, P. Zhu et al., "MimoDB 2.0: a mimotope database and beyond," *Nucleic Acids Research*, vol. 40, no. 1, pp. D271–D277, 2012.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 491612, 13 pages
http://dx.doi.org/10.1155/2013/491612

Research Article

Error Analysis of Deep Sequencing of Phage Libraries: Peptides Censored in Sequencing

Wadim L. Matochko and Ratmir Derda

Department of Chemistry and Alberta Glycomics Centre, University of Alberta, Edmonton, AB, Canada T6G 2G2

Correspondence should be addressed to Ratmir Derda; ratmir@ualberta.ca

Received 6 May 2013; Accepted 30 July 2013

Academic Editor: Yanxin Huang

Next-generation sequencing techniques empower selection of ligands from phage-display libraries because they can detect low abundant clones and quantify changes in the copy numbers of clones without excessive selection rounds. Identification of errors in deep sequencing data is the most critical step in this process because these techniques have error rates >1%. Mechanisms that yield errors in Illumina and other techniques have been proposed, but no reports to date describe error analysis in phage libraries. Our paper focuses on error analysis of 7-mer peptide libraries sequenced by Illumina method. Low theoretical complexity of this phage library, as compared to complexity of long genetic reads and genomes, allowed us to describe this library using convenient linear vector and operator framework. We describe a phage library as $N \times 1$ frequency vector $n = \|n_i\|$, where n_i is the copy number of the ith sequence and N is the theoretical diversity, that is, the total number of all possible sequences. Any manipulation to the library is an operator acting on n. Selection, amplification, or sequencing could be described as a product of a $N \times N$ matrix and a stochastic sampling operator (**Sa**). The latter is a random diagonal matrix that describes sampling of a library. In this paper, we focus on the properties of **Sa** and use them to define the sequencing operator (**Seq**). Sequencing without any bias and errors is **Seq** = **Sa** I_N, where I_N is a $N \times N$ unity matrix. Any bias in sequencing changes I_N to a nonunity matrix. We identified a diagonal censorship matrix (**CEN**), which describes elimination or statistically significant downsampling, of specific reads during the sequencing process.

1. Introduction

In vitro selection experiments—such as phage display [1, 2], RNA display, SELEX, and DNA aptamer selection [3, 4]—employ large libraries, from which 10^2–10^6 active sequences are identified through iterative rounds of selection and amplification. With the recent emergence of deep sequencing, it became possible to extract a large amount of information from the libraries before and after selection [5–10]. Deep examination of the library is a promising technique for direct evaluation of binding capacities of all binding sequences from one panning experiment. Deep sequencing also allows the characterization of unwanted phenomena in selection, such as amplification bias [6, 11].

Analysis of 10^6 reads by deep sequencing gave rise to a large number of errors that were not present in the analysis based on the small number of sequences obtained using the Sanger method. Analysis of errors in information-rich datasets is a problem with over 50 years of history; correction of digital data made of bits or words is a topic of intense research in communication theory [12]. As phage display operates with limited digital sets, data analysis techniques from the communication theory could be applied to phage display. For example, Rodi and coworkers used a positional frequency matrix to calculate the informational content or Shannon entropy of each sequence [13]. This approach could be used to distinguish potential fast growing sequences from potential hits [14]. With the introduction of deep sequencing, the problem of error analysis in phage display becomes identical to a classical information theory problem: "reproducing at one point, either exactly or approximately, a message selected at another point" [15]. The "message" is the sequence information stored in the library. Sequencing process transmits this information and makes either stochastic or predictable errors. Understanding the sources of errors during sequencing could provide mechanisms for bypassing

them, for correcting the errors, and for maximizing the amount of useful information received from sequencing.

There are over 10,000 published literature reports that contain the terms "deep sequencing" or "next generation sequencing" or any of the trademark names such as "Illumina" (reference: ISI database). Among these reports, less than 10 published reports describe sequencing of phage-displayed libraries [5–7, 9, 10, 16–19]. Deep sequencing efforts in the literature are largely focused on genome assembly and metagenomic analyses. The error analysis techniques tailored for genome assembly cannot be used directly for analysis of phage libraries because the data output from phage library sequencing is very different from the genome assembly. In genome assembly, genomic DNA is shredded into random fragments and sequenced. The genome is then assembled from these fragments *in silico*. Although multiple fragments cover each area of the genome, the probability to observe two identically shredded fragments is very small. Two exact sequences, thus, could be considered amplification artifacts and removed by error analysis software. On the contrary, in phage-display sequencing, the reads are exactly of the same length. Duplication of the same read is important for validation of the accuracy of this read. Some researchers focus exclusively on reads that have been observed multiple times and discard singleton reads as erroneous [5]. Within each library, the copy numbers of sequences range continuously by six or more orders of magnitude [5, 6, 9]. Some phage clones are observed in the entire library only a few times; other clones could be present at copy number of 100,000 per sequencing run [5, 6, 9]. Unlike multiple cells with identical genomes, each screen is unique: identical set of sequences with identical copy numbers cannot be obtained even if the screen is repeated due to stochastic number of the screen that contains low copy number of binding clones [20].

Metagenomic analyses of microorganisms recovered from environmental samples [21, 22], also known as "microbiome" [23] and "viriome" analyses [24], encountered similar problems to those observed in phage library analysis: the concentration of species observed in a particular sample is unequal [25]. The abundance of species might range by a few orders of magnitude [26]. It is possible that error analysis tools developed in the above areas could find use in phage display sequencing. For example, there are multiple published algorithms for removing errors from low copy number reads to ascertain that low copy number sequences are new species and not sequencing errors (e.g., see [27–29] and references within). Metagenomic analysis is usually more complex than analysis of phage-display libraries. First, in metagenomics, the bacterial or viral genes must be assembled from short reads *de novo*. Second, there is no simple relationship between phylogenetic classification of "species" and the observed DNA sequence. Third, the exact number of species in the environment is unknown. On the other hand, sequencing of phage-displayed peptide libraries has none of these problems: (i) it requires no assembly steps because each sequence is covered by one read; (ii) a unique DNA sequence defines a unique "species"; and (iii) the theoretical complexity in synthetic libraries is known exactly. For small libraries, such as the library of 7-mer peptides, the complexity, $(20)^7$, is within the reach of next-generation sequencing. We see phage-displayed peptide libraries as an ideal model playground for the development of optimal error analysis and error correction protocols. It is possible that error analysis developed from phage libraries analysis could then be used in other areas such as genomic and metagenomic analyses.

The errors in sequencing could be divided into "annotated" and "invisible." The "annotated" errors that originate from misincorporation of nucleotides are annotated using Phred quality score [30]. These annotated errors are removed during the processing (see below). Examples of "invisible" errors are sequence-specific frame shifts that lead to emergence of truncated reads during the Illumina sequencing [31]. Invisible errors could also originate during the preparation of the libraries for sequencing. Examples are removal of AT-rich fragments during purification of dsDNA [32] and erroneous incorporation of nucleotides during PCR [33, 34]. Mutations have the most significant impact on the observed diversity of the library. There are 63 ways to misspell a 21-mer-nucleotide sequence with a one-letter error (point mutation). The large dynamic range in concentrations of clones in the phage library exacerbates the problem. Clones that are present in high abundance—10^5 copies per read—are more prone to yield errors [6]. For example, we observed that random point mutations convert several short sequence with a copy number of 10^5 to a library of sequences with copy numbers ranging from 1 to 10^2 [11]. In attempt to unify error analysis into one convenient theoretical framework, we generalized all errors as follows. All errors either lead to disappearance of particular sequence or its conversion to another sequence of the same length. Errors, thus, operate within a finite sequence space, and it should be possible to use elementary linear algebra to generalize most processes that lead to errors.

2. Theoretical Description

See Table 1.

2.1. Operator Description of the Phage-Display Library and Selection Process. In our previous reports, we described the phage library as a multiset, or a set in which members can appear more than once [35]. This description also simplifies the analysis of the errors in these libraries. The multiset description represents a library with N theoretical members as an ordered set of N sequences and N × 1 copy number vector (n) with positive integer copy numbers (Figure 1(a)). Any manipulation of a phage library—such as erroneous reading or selection—changes the numbers within the copy number vector. All manipulations to the multiset, thus, could be described by operators (**Op**) that convert vector n_1 to another vector n_2 as $n_2 = \mathbf{Op}\, n_1$ (Figure 1(c)). For an N × 1 vector, the operator is N×N matrix. If elements are selected or eliminated independently of one another, the N × N matrix is diagonal (Figure 1(d)). This approach is uniquely convenient for libraries of short reads. For example, a library of 7-mers contains exactly $20^7 = 1.28 \times 10^9$ peptides and is described

Table 1: Symbols and definitions used in the theoretical description section.

Symbols	Meaning
A, a, f, m, n, k	Unless specified otherwise, normal font designates scalars
$A, a, N, P, {}^{1}\mathrm{n}, {}^{13}\mathrm{n}$	Italic font designates vectors. Different vectors can be distinguished by the left-superscript notation
A, a, Abc, Pan, Sa	Bold font designates operators or matrices (here all operators are matrices)
$^{1}\mathbf{A}, {}^{\mathrm{f}}\mathbf{Sa}, {}^{0.9}\mathbf{Sa}, {}^{0.5}\mathbf{Sa}$	Operators can be distinguished by the left-superscript notation. For sampling operator **Sa**, this notation specifies the sampling fraction of the **Sa** operator
$\mathrm{A}_1, \mathrm{a}_2, \mathrm{A}_i, \mathrm{a}_j$	Normal font with right subscript designates scalar values of the vector
$\mathrm{A}_{11}, \mathrm{A}_{21}, \mathrm{A}_{ij}, \mathrm{A}_{ii}$	Normal font with two right subscripts designates scalar values of the 2D matrix
$\lVert \mathrm{A}_1 \dots \mathrm{A}_5 \rVert$	Description of the scalar elements in the vector
$\lVert \mathrm{A}_{ij} \dots \mathrm{A}_{ii} \rVert$	Description of the scalar elements in the matrix
$\mathrm{x} \in [\mathrm{A}\ \mathrm{B}]$	Scalar x belongs to the inclusive scalar interval [A B]; that is, $\mathrm{A} \le \mathrm{x} \le \mathrm{B}$
$x \in [A\ B]$	Vector x belongs to the "vector interval" $[A\ B]$; that is, for every element $\mathrm{A}_i \le \mathrm{x}_i \le \mathrm{B}_i$
{A B C . . . X}	Set where A, B, C,. . ., X are the unique elements of the set
{A(a) B(b). . .X(x)}	Multiset (2-tuple) where A, B,. . ., X are the unique elements and a, b, x are the scalars describing the copy numbers of the A, B, X elements
I_N	Unity matrix of the Nth order; that is, N × N matrix $\lVert \mathrm{A}_{ij} \rVert$, $\mathrm{A}_{ij} = \delta_{ij}$ (Kronecker delta)

completely using a 10^9-element vector. This size is accessible to the computational capacity of most desktop computers.

In operator notation, phage display can be described as

$$\mathrm{Sel} = \mathbf{Pan}\ \mathrm{Naive}, \tag{1}$$

where *Naive* is the copy number vector for naïve library, *Sel* is the copy number vector after panning, and **Pan** is a panning operator. In standard phage display, the **Pan** operator is a complex product of all manipulation steps (binding, amplification, dilutions, etc.). If a screen uses no amplification and uses deep-sequencing [9, 16], or large-scale Sanger sequencing [36, 37] to analyze the enrichment, it might be possible to define the panning process as a simple product of two operators as follows:

$$\mathbf{Pan} = {}^{\mathrm{f}}\mathbf{Sa}\ \mathbf{K}_\mathbf{a}, \tag{2}$$

$$\mathrm{Sel} = {}^{\mathrm{f}}\mathbf{Sa}\ \mathbf{K}_\mathbf{a}\ \mathrm{Naive}, \tag{3}$$

where $\mathbf{K}_\mathbf{a}$ is a deterministic "association" operator, which contains association constants for every phage clone present in the library. Description of such operator is beyond the scope of this paper and we recommend consulting other reports that attempted to generalize the selection procedure [20]. Another operator in (3) is a sampling operator ($^{\mathrm{f}}\mathbf{Sa}$), which describes stochastic sampling of the library with m sequences to yield a sublibrary with f^* m-sequences, where $\mathrm{f} \in [0\ \ 1]$ is a sampling fraction. $^{\mathrm{f}}\mathbf{Sa}$ operator has the following properties, which emanate from physical properties of the sampling procedure:

(I)

$$^{\mathrm{f}}\mathbf{Sa}_\mathrm{i}\ 0 = 0\ (\text{sampling does not create new sequences from nonexisting sequences}). \tag{4}$$

(II) $^{\mathrm{f}}\mathbf{Sa}$ is a diagonal operator with diagonal scalar functions $\lVert \mathrm{Sa}_{11}\ \mathrm{Sa}_{22} \cdots\ \mathrm{Sa}_{\mathrm{NN}} \rVert$, $\mathrm{Sa}_\mathrm{i}(0) = 0$.

(III) In $B = {}^{\mathrm{f}}\mathbf{Sa}\ A$, B is a vector of positive integers, $\mathrm{B}_i \ge 0$ and $\mathrm{sum}(B) = \mathrm{f}^*\mathrm{sum}(A)$. Integer values ensure that the observable values of the operator have physical meaning. The clone could be observed once (1), multiple times (2, 3, etc.), or not observed at all (0).

(IV) **Sa** is nondeterministic operator. When applied to the same vector, the operator does not yield the same result but one of the possible vectors that satisfy rules (I–III). The majority of the solutions of the operator, however, reside within a deterministic confidence interval $^{\mathrm{f}}\mathbf{Sa}\ A \in [{}^{\mathrm{lo}}C\ \ {}^{\mathrm{hi}}C]$.

(V) As a consequence from (IV), operator **Sa** is nonlinear, noncommutative, and nondistributive.

(VI) Large sum of sampling operators with same f should "average out" to yield I_N unity matrix

$$\frac{\left({}^{\mathrm{f}}\mathbf{Sa}_1 + {}^{\mathrm{f}}\mathbf{Sa}_2 + {}^{\mathrm{f}}\mathbf{Sa}_3 + \cdots {}^{\mathrm{f}}\mathbf{Sa}_\mathrm{k}\right)}{\mathrm{k}} \longrightarrow \mathrm{f}^*\mathrm{I}_\mathrm{N}, \quad \text{as } \mathrm{k} \longrightarrow \infty. \tag{5}$$

The **Sa** operator is simple to implement as a random array indexing function in any programming language (e.g., see Supplementary Schemes S1, and S2 available online at http://dx.doi.org/10.1155/2013/491612). It might be possible to express $^{\mathrm{f}}\mathbf{Sa}$ analytically for any f as a diagonal matrix (Figure 1(d)). In this paper, we use numerical treatment by an array sampling function because it is more convenient for multisets of general structure. We tested the random indexing implementation to show that the sampling algorithm yields a normal distribution for a large number of samples (Supplementary Figure S1). Despite the simplicity of $^{\mathrm{f}}\mathbf{Sa}$

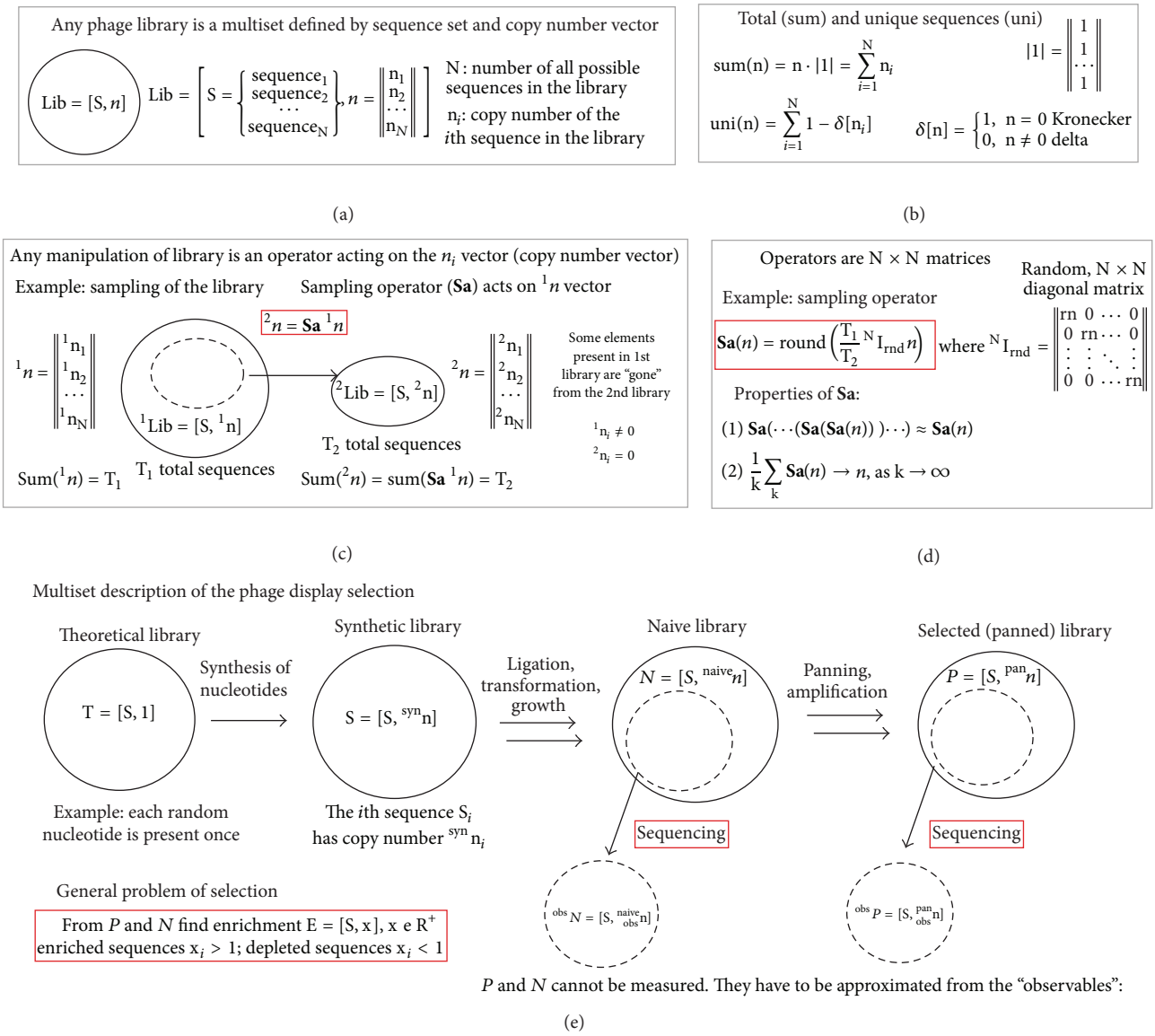


FIGURE 1: (a) Phage library can be described by multisets made of S = {sequence set} and n = ‖vector of copy numbers‖. Any change to the library can be described as function/operator acting on the n. (b) Relevant functions are calculations of total sequences (sum) and unique sequences (uni). (c) Any transformation of library to another library is an operator acting on n. Sampling of libraries to yield a sublibrary is the most important operator. (d) It can be described as N × N matrix. Specifically, **Sa** is a diagonal matrix of values derived from random distribution. Rounding function is necessary to ensure the physical meaning of the sampling results. **Sa** acting on the same vector yields one of many vectors that have the same number of total elements. As a consequence, **Sa** is nonlinear, nondistributive, and noncommutative operator. Average of many **Sa** operators is a scalar (dilution factor). (e) Any screen of any library can be described as operators acting on the copy number vectors of the naïve (or theoretical) library. Copy number vectors cannot be observed directly. They have to be measured through sequencing. As sequencing contains sampling process (**Sa** operator), the result of sequencing is nondeterministic. Sequencing yields one of many possible *observed* copy number vectors, none of which are equal to the *real* copy number vector.

implementation—the entire code is <30 lines in MatLab—the script allows rapid calculation of the results of ^{f}**Sa** for a multiset of reasonable size (several million sequences, Figures 4 and 5) on a desktop computer.

We evaluated the behaviors of $^{0.5}$**Sa** for several multisets. The probability to observe a specific solution is described in Figure 3(b). Individual solutions can be represented as lines with nodes on XY-plane, where each node represents one element of the multiset (Figures 3(d) and 3(e)). The most probable solutions reside near the "expected solution" (represented as dotted line), and the probability to observe a solution where many elements deviate from the probable solution is low (Figure 3(e)). Graphical representation of the solutions highlights that sampling could lead to deviation of the frequency of the individual elements of the multiset; for example, Figure 3(e) describes >2 fold deviation from

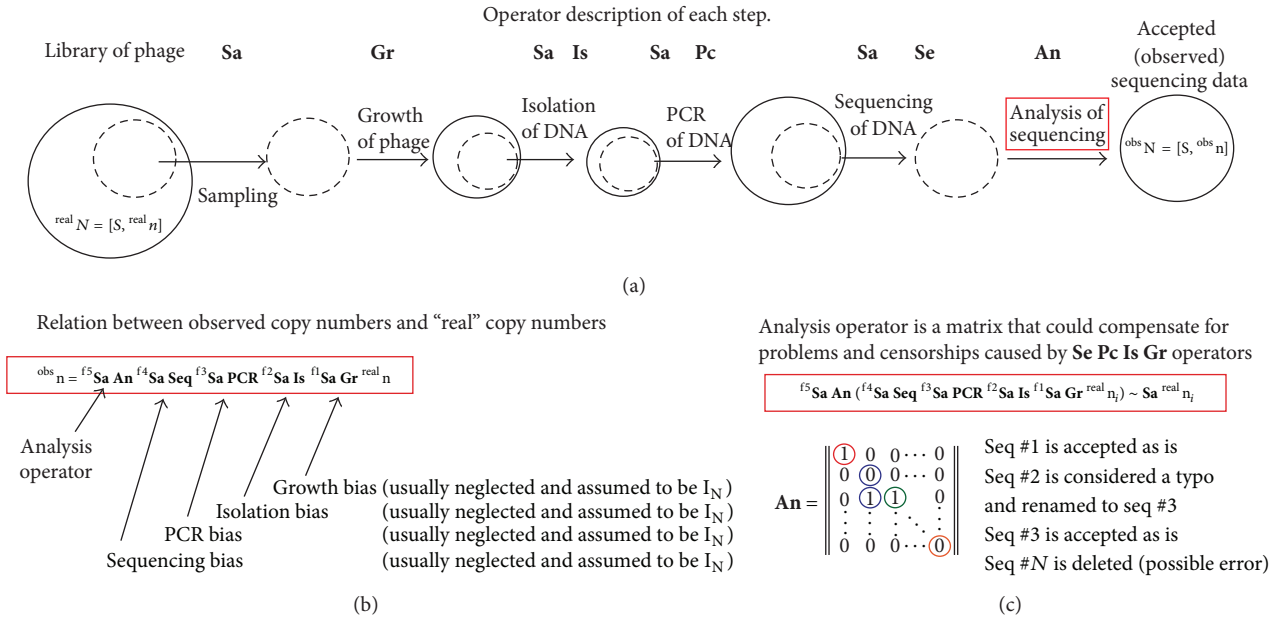


Figure 2: Operator description of the deep sequencing process. (a) A library of phage must be processed before deep sequencing. Each step involves sampling, which is either a deliberate partitioning of the sample or random loss of the sample. Each sample preparation state could (and does) introduce bias in sequence abundance. Each step, thus, is an operator chat changeing the n vector. (b) If we ignore bias during preparation, operators could be approximated as unity vectors, and sequencing could be represented as a product of sampling and analysis operators. (c) Analysis operator (**An**) is a binary decision matrix, which describes what sequences are and are not considered as errors. Decisions, such as removal of sequences or correction of sequences, are the most important because they decide which "*observed*" sequences are considered "*real*." To make the analysis of the selection process meaningful, the same **An** operator should be used in all analyses.

the expected value for one of the elements. Figure 3(f) shows that the solution in which two elements deviate by >2 fold is improbable. This observation is a simple consequence of the multiplicity of the probabilities (large deviation from the average has probability p and the probability to observe this deviation twice is p^2).

Even in small multisets, such as {A(1) B(2) C(3) D(4)} made of four unique and 10 total elements, $^{0.5}$**Sa** {A(1) B(2) C(3) D(4)} operation yields large number of solutions with equal probability, termed as redundant solutions (e.g., solutions that have equal probability in Figure 3(b)). Redundancy depends on the structure of the multiset (Figure S2). This redundancy makes the calculations of all probable solutions of **Sa** impractical. For sets even with 5-6 unique elements, identification of all vectors B, which satisfy equation $B = {}^{f}\mathbf{Sa}\ A$ and reside within a 95% interval, requires hundreds of thousands of iterations (Figure S2 and S3). On the other hand, calculation of the confidence interval of each element B_i of the vector B converges rapidly. A multiset $\{A_{1000}\} = \{A_1(1)\ A_2(2) \cdots A_{1000}(1000)\}$ with 1000 unique elements and $1 + 2 + 3 + \cdots + 1000 = 500,500$ total elements is similar to an average deep sequencing data set (Figure 4). Calculation of all probable solutions of $^{0.5}$**Sa** $\{A_{1000}\}$ is beyond the capabilities of most computers. However, the 99.9% confidence interval of all elements of vector $B = {}^{0.5}\mathbf{Sa}\ \{A_{1000}\}$ can be calculated in ~2 minutes on an average desktop computer. The red dots in Figure 4 are $^{lo}C_i$ and $^{hi}C_i$ or the 99.9% high and low confidence interval of all elements B_i (Figure 4).

The sampling operator is critical in phage display because sampling of libraries occurs in every step of the selection and the preparation of libraries for sequencing. The stochastic nature of sampling operators makes two identical screens "similar within a confidence interval." Solving (1) exactly is not possible, but it should be possible to estimate the solution within a confidence interval. Consider

$$\mathrm{Sel} \in \left[{}^{lo}\mathbf{K_a}\ \mathrm{Naive};\ {}^{hi}\mathbf{K_a}\ \mathrm{Naive} \right], \tag{6}$$

where $^{lo}\mathbf{K_a}$ and $^{hi}\mathbf{K_a}$ are diagonal matrices of the upper and lower confidence intervals for the association constants. A simulation of the behavior of the **Sa** operator (Figures 3 and S3) suggests that the relative sizes of the confidence intervals might be impractically large when the copy numbers of sequences are <10.

Multiple sampling events of the **Sa** operator yield a normal distribution for each element of the vector (Figure 3). Fitting this normal distribution could yield a "true" value of the process. This process is identical to the extrapolation of the average from the normal distribution of noisy data. Multiple algorithms for such extrapolation exist for one- and multidimensional stochastic processes [38, 39]. We believe that **Sa** behaves as a one-dimensional stochastic process and it might be possible to extrapolate the true value of the sampling from 7 to 10 repeated instances of **Sa** (i.e., the number of data sufficient to fit an 1D normal distribution). The necessary practical steps towards solving (3) or (10) are the following. (i) Eliminate or account for any bias not related to binding (e.g., growth bias). (ii) Repeat the screen

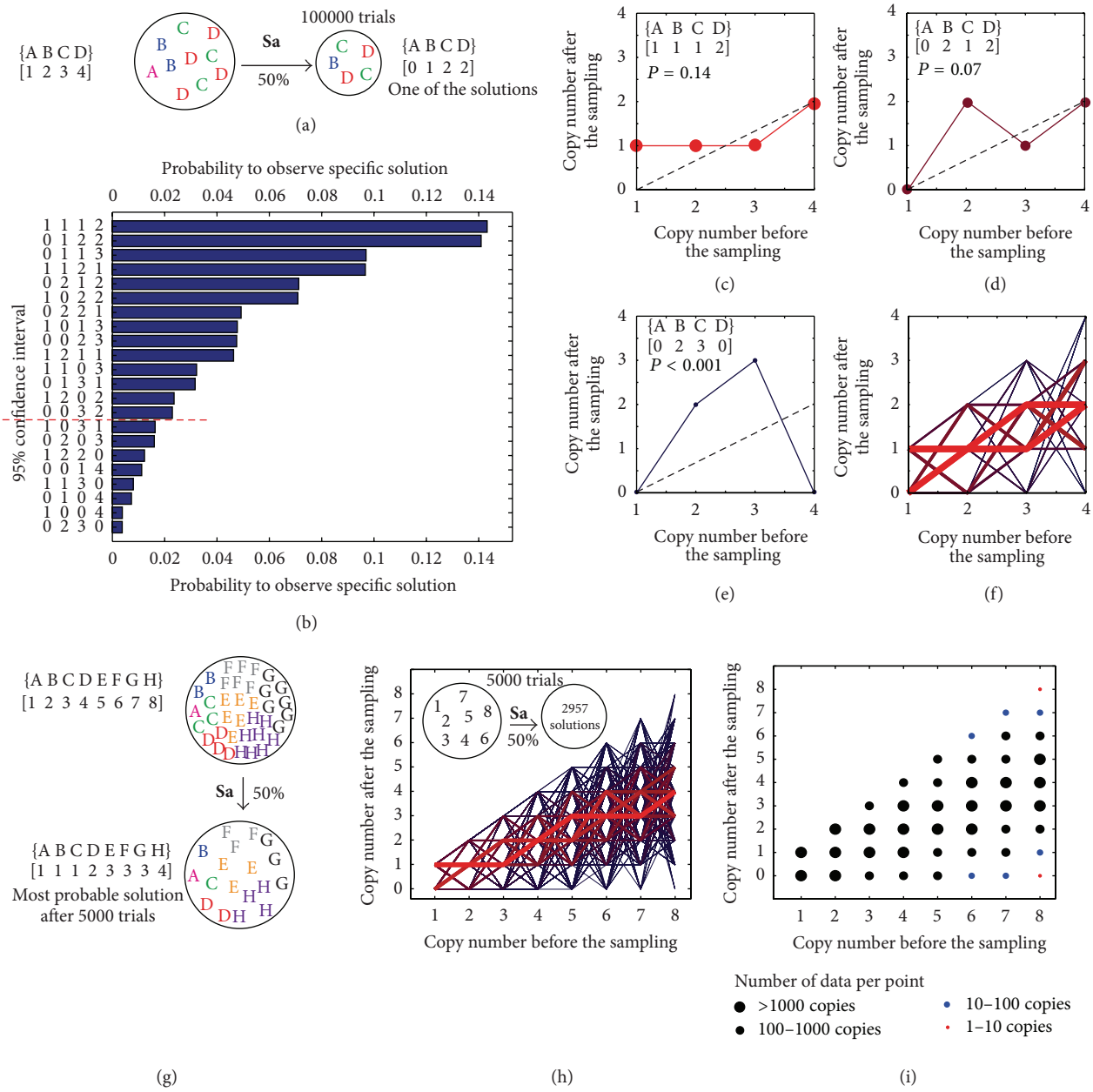


FIGURE 3: (a) Testing the sampling operator implemented as random indexing function using a model multiset. (b) In 100,000 trials, we observed 22 unique solutions from which 14 resided in a 95% confidence interval. Solutions with 0 and 1 copies of element A were found at equal abundances ("redundant solutions"). (c) Representation of the most probable solution as a line with 4 nodes; "p" is a probability to find the solution; dotted line is an expected "average solution" for 50% sampling. (d) The 5th most probable solution; (e) least probable solution deviates the most form the average; (f) combination of all solutions. Red thick lines describe the most probable solutions; thin blue lines describe the least probable solutions. (g) Sampling of larger multisets yields more possible solutions (here, 2957 in 5000 trials). (h) All solutions of the sampling represented as lines. (i) Probability to observe a particular copy number after sampling. While (h) is the most accurate representations of the confidence intervals, the thin blue lines describe solutions outside the confidence interval; this representation is impractical due to large number of redundant solutions in larger multisets. In (i), confidence interval could be extrapolated from distributions of individual copy numbers (e): red dots are on or outside the confidence interval.

several times. (iii) Measure all copy numbers of all sequences, including zero values, with high confidence. Requirement (i) has been an ongoing effort in our group [11, 40] and other groups [13, 41–43]; for review see [11, 44]. Deep sequencing makes it simple to satisfy requirement (ii) and obtain multiple instances of the same experiment. For example, we described the Illumina sequencing method that allows using barcoded primers to sequence 18 unrelated experiments in one deep

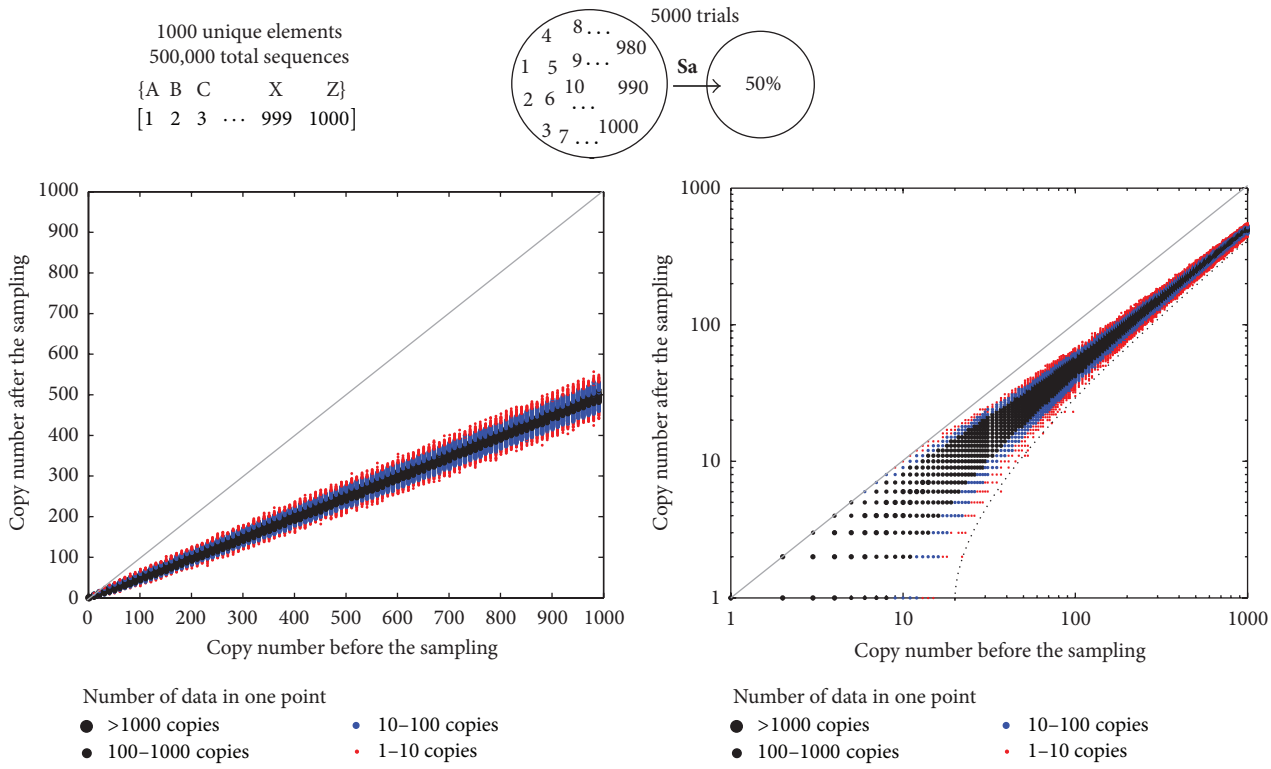


Figure 4: (a) Testing the sampling operator using a large multiset made of 1000 unique elements with 1000 different copy numbers. Images describe linear and log-scale representation of the confidence interval of the sampling operator. Solutions beyond this interval were not observed in 5000 trials. Dotted line represents an overestimate of the 99.9% confidence interval (for details, see Figure S4). Most probable outcomes of the **Sa** operator have either zero or one unique sequence beyond this interval. This line is used in subsequent sections (Figures 5 and 6). We note that distributions of the copy numbers have well-defined shape; according to central limit theorem, it is a normal distribution. With enough replicas, it should be possible to extrapolate the center of this distribution, define the solutions explicitly, and bypass the stochastic nature of the **Sa** operator.

sequencing experiment [45]. We recently scaled this effort to 50 primer sets and evaluated the performance replicas of simple selection procedures (in preparation).

The measurement of the copy numbers of sequences is a separate problem that can be described using the same sampling operators and bias operators that describe how the library is skewed by each preparation step. For example, isolation of DNA by gel purification disfavors AT-rich sequences, whereas PCR favors sequence with within specific GC-content range [32]. The *real* sequence abundance in any phage library ($^{\text{real}}n$), hence, has to be derived from the *observed* sequence abundance ($^{\text{obs}}n$) by solving this equation:

$$^{\text{obs}}n = \left(^{\text{f5}}\mathbf{Sa\ An}\right)\left(^{\text{f4}}\mathbf{Sa\ Seq}\right)\left(^{\text{f3}}\mathbf{Sa\ PCR}\right) \times \left(^{\text{f2}}\mathbf{Sa\ Is}\right)\left(^{\text{f1}}\mathbf{Sa\ Gr}\right)\ ^{\text{real}}n. \quad (7)$$

In this equation, each operator in brackets describes a bias at a particular step. $^{\text{f}}$**Sa** describes sampling at that step, and f1–f5 describe the sampling fractions. The bias in growth (**Gr**), isolation (**Is**), PCR amplification (**PCR**), and sequencing (**Seq**) could be related to the nucleotide sequences. The **An** analysis operator is a matrix that describes retaining, discarding, or correcting the sequence (Figure 2(b)). An ideal **An** operator could compensate for the biases introduced by another operator (Figure 2(c)). To define such operator, (7) could be potentially solved using repeated sequencing of a well-defined model library. In the next applied section, we examine the real deep sequencing data and identify conditions under which these operators could be at least partially defined.

2.2. Analysis or the Error Cutoff in Deep Sequencing Reads. All next-generation sequencing techniques provide quality score (Phred Score) for every sequenced nucleotide. In Illumina sequencing, this score is related to the probability of the nucleotide being correct [46]. In low throughput Sanger sequencing, the Phred score monotonously decreases with read length and the mechanisms that yield errors in capillary electrophoresis are well understood. Common practice in Sanger sequencing is to discard all reads after the first nucleotide with a Phred score of 0. In next-generation sequencing, the filtering of the reads is usually more stringent as follows.

(A) Discard reads that have at least one read that has score lower than "cutoff."

(B) Discard reads that had cumulative Phred score lower than cutoff.

Tags			NKKN	Bar	Adapter	Sequence	End of read			Quality string for sequence	
C7C	<F>	<PERF>	TGGG	CGA	TATTCTCACTCT	GCTTGTCCTCCGAAGGTGCATCTGCAGTGCGGT	~~~~~~~	CCCF	FFF	FHHHHHJJJJJJ	JJJJJJJJJJJJJJJHHHGIJJJJJ>FHJJIJH~~~~~~~
C7C	<F>	<PERF>	GGGC	TTG	TATTCTCACTCT	GCTTGTGCGGCGAATTCTTATCGGAATTGCGGT	~~~~~~~	CCCF	FFF	FHHHHHJJJJJJ	JJJJJJJJJJJIIJJJJJJIJIJJIEHHHFHFD~~~~~~~
Ph7	<F>	<PERF>	TGTG	GAC	TATTCTCACTCT	ACTCAGCCTCATCATACTCCGGGTGGAGGTTCG	~~~~~~~	CCCF	FFF	FHHHHHJJJJJJ	JJJJJJJJJJJJJJJJJJJJJJHIJJJJJJJJ~~~~~~~
Ph7	<F>	<PERF>	ATGG	GAC	TATTCTCACTCT	GCGACGACTGTTCCAGCTTCGGGTGGAGGTTCG	~~~~~~~	CCCF	FFF	FHHHHHJJJJJJ	JJJJJJJJJHJJIJHIIJJJBFHAHIHHHF~~~~~~~
C7C	<F>	<PERF>	TTGG	GTA	TATTCTCACTCT	GCTTGTGATGCGCGTATGAATCAGCCTTGCGGT	~~~~~~~	CCCF	FDF	FHHHHHJJJJJJ	JJJJJJJJJJJIIJJJJJJJJJJJJJJJJJH~~~~~~~
C7C	<F>	<PERF>	CGGC	TTG	TATTCTCACTCT	GCTTGTCAGGATAAGAATTCGCTTTTTTGCGGT	~~~~~~~	@CCF	FFA	DFFHBFFFGEFH	GEGIDFAEGI@FCHC9F@F?@@AGH*?D?BAH; ~~~~~~~
C7C	<F>	<PERF>	CGGG	CGA	TATTCTCACTCT	GCTTGTAAGTCTCTTTCGAGTGCTCTGTGCGGT	~~~~~~~	CCCF	FFF	FHHHHHJJJJJJ	JJJJJJJJJJJJJJJIGHGIHIIIGHGIJJJJH~~~~~~~
C7C	<F>	<PERF>	CGGG	TTG	TATTCTCACTCT	GCTTGTCCTCGTCTGTATCAGATGTCGTGCGGT	~~~~~~~	?@@F	=A:	ADDDHFEFIEHG	GAHGHFHGGDH:EHB?GGCEHIEGB@FDHBD<A~~~~~~~
Ph7	<F>	<PERF>	CGTC	TCA	TATTCTCACTCT	GATGCGCATCGTACGTGTAATGGTGGAGGTTCG	~~~~~~~	CCCF	FFF	FHGGHHJJJJJJ	JJIJJIGIJIJEGIJGHFGIIJJ?FFGGIIGII~~~~~~~
Ph7	<F>	<PERF>	TGTG	TCA	TATTCTCACTCT	CGGGCTGCGTGTGCGAATCCTGGTGGAGGTTCG	~~~~~~~	CCBF	FFF	FHHHHHJJJJJJ	JJJJJJJJJJJJGGHIJJJJJJJIJJIJJJJJJ~~~~~~~
P12	<F>	<PERF>	ATGC	CGA	TATTCTCACTCT	TGGCGTCATGATTCGTTTTAGACGGTTCAGCGT	~~~~~~~	CCCF	FFF	FHHHHHJJJJJJ	JJJJJJJJJJJIJJIJJJJJJJJJJJJIJJJ~~~~~~~

(a)

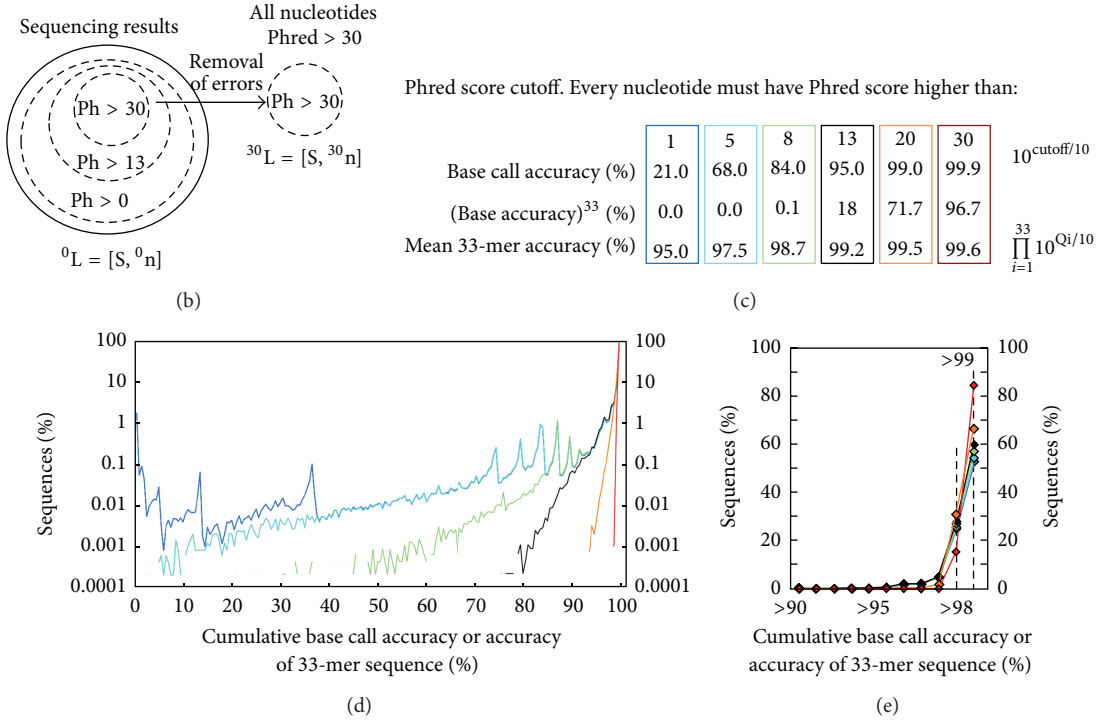


Figure 5: (a) Representative lines from the intermediate file from Illumina deep sequencing analysis (for more information see ERROR_TAG_data0001.txt in Section 4 and our previous publication [6]). The reads have been parsed to identify adapters and barcodes. Each read has been tagged according to the library type, direction of the read, and quality of the adapter regions. We use this intermediate library to identify reads that harbor erroneous nucleotides. (b) Multiset view of the intermediate library. The library contains subsets that have low, medium, and high quality reads. Error filtering of this intermediate library to eliminates any read with Phred score below 30 yields a high quality library of reads ^{30}L. (c) Mean accuracy of the reads in the library after error filtering ranges from 95% to 99.6%. Even for very low-quality cutoff, Phred > 1, the average read quality is 95%. (d) Distribution of cumulative read accuracy in libraries processed using different cutoffs. (e) Linear plot of the data presented in (d) with zoom in on the region with >90% cumulative accuracy.

(C) Use a combination of A and B (accept reads with minimal cutoff and minimal cumulative score).

Many of the error analyses in the area of deep sequencing are designed for genetic reads, which have variable lengths and unknown sequence throughout the whole read. Analysis of the reads in a phage-display library is a simpler problem because phage-derived constant adapter regions flank the variable reads. Identification of the adapter region is a necessary first step in the analysis. Reads, in which the adapters cannot be mapped, cannot be used. We designed algorithms to recover reads, even if adapters were hampered by truncation, deletion, or mutation [6]. We observed that the reads flanked by the erroneous adapters had a significantly higher error rate than reads flanked by "perfect" adapters (unpublished). Example of mapping of flawed reads in Illumina sequencing is provided in ERROR_TAG_data0001.txt (see Section 4). In the remaining sections, we analyze the population of the sequences preceded by a "perfect" adapter to identify possible sequence-specific biases.

We analyzed a typical library sequenced by Illumina using various cutoffs (Figure 5). We analyzed a 33-bp segment of the library that contained variable seven amino acids and a constant region and GGGS terminus. A simple cutoff that discards reads with Phred <1 nucleotides yields library

termed ${}^{1}n$, which had an average 95% accuracy of the 33-nucleotide read. Reads that do not contain Phred = 0 nucleotide rarely contain multiple low-quality reads. The ${}^{1}n$ library was bimodal: 80% of the reads had overall accuracy of 99%, very few reads with accuracy 5–90%, and significant number of reads with accuracy of 1% (Figures 5(d) and 5(e)). These observations suggest that reads can be divided into (i) reads free of errors and (ii) reads with multiple errors.

An example of a more stringent cutoff is elimination of reads with Phred <13 nucleotides; this process yielded a library ${}^{13}n$ in which every nucleotide had >95% confidence. The number of total reads in ${}^{13}n$ was 10% less than number of reads in ${}^{1}n$; that is, sum($ {}^{13}n$) = 0.9sum($ {}^{1}n$). The observed average read accuracy of the read in the ${}^{13}n$ library was 99.2%. Theoretically, the 0.95 confidence cutoff in a 33-mer nucleotide could yield reads with accuracy as low as $(0.95)^{33} = 18\%$. In practice, the probability to find reads with multiple nucleotides of 95% accuracy was vanishingly small. Specifically, among 500,000 reads, the lowest observed cumulative accuracy was 77%. Such a result, for example, could be obtained in a sequence that has 27 "perfect" nucleotides and 5 nucleotides with a Phred = 13 score: $(1)^{27}(0.95)^{5} = 0.77$. Applying the most stringent cutoff to eliminate all reads with a Phred < 30 yielded a library ${}^{30}n$ in which every nucleotide had 99.9% confidence. The average confidence of the reads improved subtly from 99.2% to 99.6%. The number of total reads in ${}^{30}n$ was 30% less than number of reads in ${}^{13}n$; that is, sum($ {}^{30}n$) = 0.7sum($ {}^{13}n$). It was not clear whether such cutoff is an improvement or a detriment for analysis. In the next section, we examined how frequency of the members of the library changed upon application of each error cutoff.

2.3. Example of Error Analysis: Sequence-Specific Censorship during Phred Quality Cutoff. If errors occur by random chance, they should be uniformly distributed in all sequences. Removal of erroneous read, in that case, should be identical to sampling of the library by ${}^{f}\mathbf{Sa}$ operator, where f is the sampling fraction. For example, consider the removal of Phred < 13 nucleotides from an unfiltered library (process denoted as ${}^{1}n \rightarrow {}^{13}n$). From the experiments, we know that sum($ {}^{13}n$) = 0.9sum($ {}^{1}n$); if errors were distributed in sequences at random, the ${}^{1}n$ and ${}^{13}n$ vectors should be related as

$$ {}^{13}n = {}^{0.9}\mathbf{Sa}\left({}^{1}n\right). \tag{8}$$

The solutions should reside within a confidence interval

$$ {}^{13}n \in \left[{}^{\text{lo}}C \quad {}^{\text{hi}}C\right]. \tag{9}$$

If errors occur preferentially in specific reads, the frequency of these reads should occur beyond the confidence interval of the ${}^{0.9}\mathbf{Sa}$. This process could be described by a diagonal matrix **Bias** as

$$ {}^{13}n = {}^{0.9}\mathbf{Sa}\left(\mathbf{Bias}\left({}^{1}n\right)\right). \tag{10}$$

The elements of the diagonal matrix $\mathbf{Bias} = \|\mathrm{B}_{ii}\|$ could be estimated as follows:

$$ {}^{13}n_i \in \left[{}^{\text{lo}}C_i \quad {}^{\text{hi}}C_i\right], \qquad \mathrm{B}_{ii} = 1, \tag{11}$$

$$ {}^{13}n_i < {}^{\text{lo}}C_i, \qquad \mathrm{B}_{ii} = \frac{{}^{13}n_i}{\left(0.9\,{}^{1}n_i\right)}. \tag{12}$$

Figure 6(c) describes the representative solution of the ${}^{0.9}\mathbf{Sa}({}^{1}n)$ (green dots) and the confidence interval (blue lines). Supplementary Scheme S3 describes the script that calculated this interval from multiset ${}^{1}n$, described as a plain text file PhD7-Amp-0F.txt, using 10,000 iterative calculations of ${}^{0.9}\mathbf{Sa}({}^{1}n)$. This calculation required ~2 hours on a desktop computer. Confidence interval was estimated as the minimum and maximum copy number found after 10,000 iterations. In this approximation of the confidence interval, for sequences with the copy number <10 before sampling, it was impossible to determine whether the sequence disappeared due to random sampling or due to bias. The values of **Bias** operator cannot be defined for these sequences and it could be assumed to be 1 (see (11)). For copy number >10, however, sequence-specific bias can be readily detected. We observed that the removal of Phred <13 reads yielded a multiset in which a large number of sequences deviated beyond the confidence interval (Figure 6(d)). Their sequences could be readily extracted by comparing the vector ${}^{13}n$ with the vector of the lower confidence intervals ${}^{\text{lo}}C$ (see (12)). The solution of the **Bias** can be illustrated graphically (Figure 6(e)). Top 30 censored sequences are listed in Table S1; the other sequences can be found in the supplementary information (file PhD7-Amp-0F-13F-CEN.txt).

We performed similar calculations for ${}^{1}n \rightarrow {}^{30}n$ and ${}^{13}n \rightarrow {}^{30}n$ processes. The latter process is the most interesting because ${}^{13}n$ library has all nucleotides within acceptable confidence range (>95%) and the distribution of cumulative quality suggested that errors, on average, do not cluster in one read (Figure 5). The ${}^{13}n \rightarrow {}^{30}n$ conversion eliminated 30% of the reads, and copy numbers of many sequences deviated significantly from the random sampling; these sequences are represented by green dots outside the blue confidence interval in Figure 6(h). Top 30 sequences are listed in Table S2. The censorship is not only sequence-specific, but also position-specific. In sequences that had been censored during the ${}^{13}n \rightarrow {}^{30}n$ process, lower quality reads clustered around 3-4 specific nucleotides (supplementary information Figure S5).

The mechanism that leads to the disappearance of censored sequences is not currently clear. We attempted to identify common motifs in censored sequences using two approaches: (i) clustering and principal component analyses based on Jukes-Cantor distance between sequences and (ii) identification of motifs using multiple unique sequence identifier software (MUSI) [17]. These approaches could not detect any property common to censored reads, which would make them significantly different from the other, noncensored reads. Still, we hypothesize that the observed

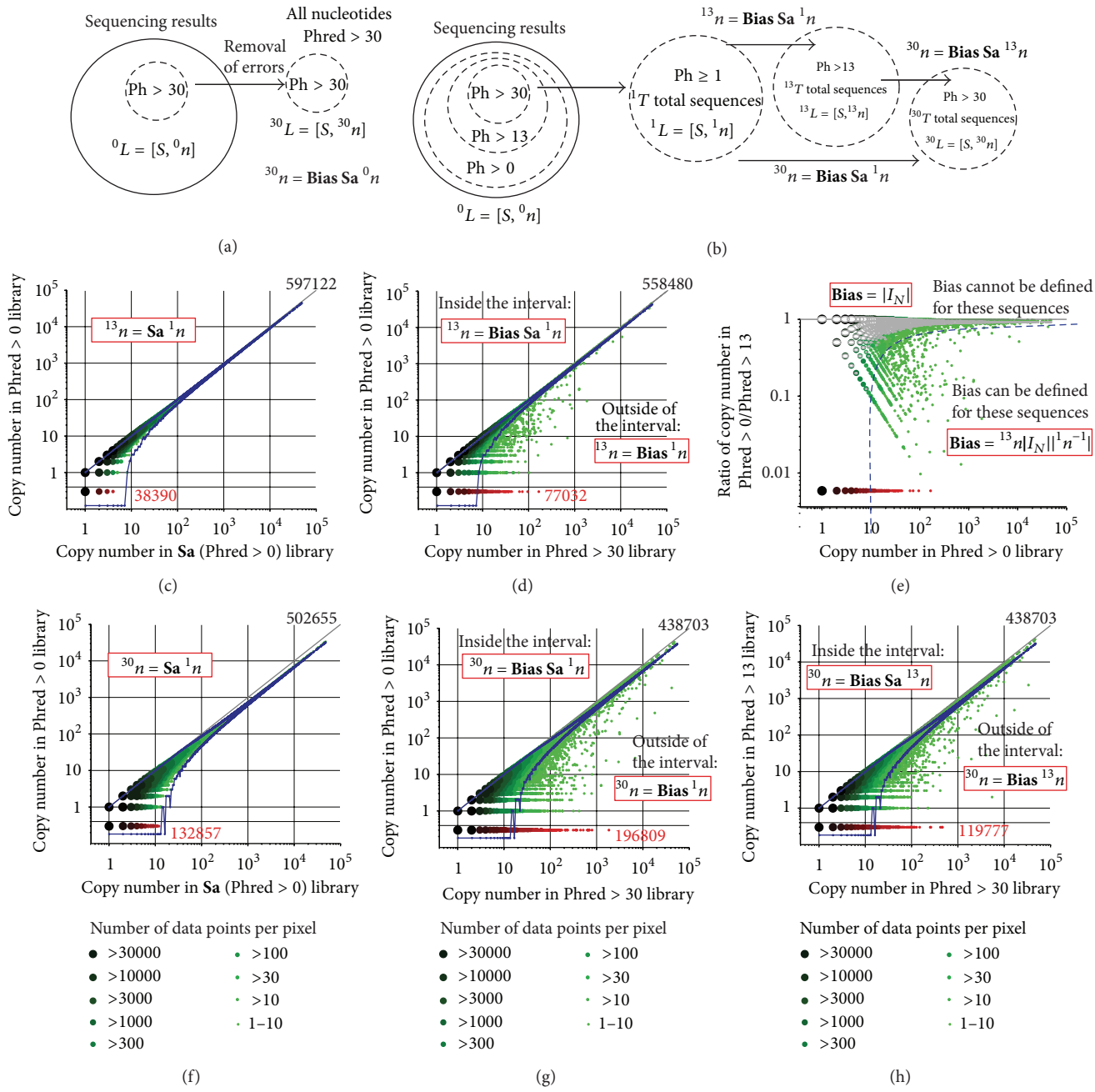


Figure 6: (a) Operator and multiset description of the error filtering procedure. Applying a Phred > 30 cutoff to library filtered by Phred>1 cutoff (^{1}n) yields a subpopulation of the library (^{30}n). If errors are sequence-independent, the ^{1}n → ^{30}n process should be identical to random sampling (^{30}n = **Sa** ^{1}n). Any sequence-specific bias (**Bias**) should be detected as deviation from **Sa** ^{1}n. (b) Progressive sampling with more stringent cutoff. (c) Theoretical **Sa** ^{1}n and theoretical 99.9% confidence interval (blue). (d) Observation of statistically significant deviation from **Sa** operator: dots beyond the blue line represent sequences prone to bias. Red dots represent sequences that disappeared after in ^{1}n → ^{30}n process or during **Sa** ^{1}n sampling. (e) Magnitude of the bias range from 5 to 100-fold. (f) Bias in sampling of Phred > 30 data from Phred > 1 data ((f) is theory, (g) is observed). (h) Bias upon sampling of Phred > 30 data from Phred > 13. Many sequences were lost in this sampling and this loss was statistically significant beyond the 99.9% interval. This result shows that some sequences have propensity to harbor low- and medium-quality reads. Distribution of the errors is sequence specific.

censorship represents sequence-specific errors, which occur in every time such sequence passes though the Illumina analyzer. For example, the sequences listed in Tables S1 and S2 and supplementary files were censored in five independent experiments, which were pooled and processed simultaneously in one Illumina run. Analysis of other instances of Illumina sequencing performed by other groups could help prove (or disprove) that censorship is indeed sequence-specific

and experiment-independent. Sequence-specific censorship during Illumina analysis has been described in other publications [46]. The observations presented above suggest that reading of some sequences in phage libraries does not yield an accurate copy number. Even if these sequences were enriched due to binding, their apparent copy number in sequencing would be decreased due to sequencing bias. If the magnitude of bias is known, however, such error could be corrected. We anticipate that other biases could be calculated for these and other libraries in similar fashion. Their calculation extends beyond the scope of this paper and it will be performed in our next publication.

3. Discussion

3.1. Significance and Transformative Potential of Library-Wide Error Correction. In the Medicinal Chemistry field, structure-activity relationships (SAR) and pharmacophores are built using both positive and negative observations. It is the negative results that bear the most significance in these studies because they allow mapping of the range of conditions under which particular structure no longer works. For example, SAR of an R group of a ligand might be built on the following observations. A ligand binds to the target when the R group in the specific position is methyl or ethyl; changing R to *iso*-propyl and *tert*-butyl ablates the binding. This concludes that the R group must be a small alkyl group. An analogous situation is found in SAR of peptide ligands; the most important information from alanine scan mutagenesis is loss of function because it helps identifying the important residues. Interestingly, loss-of-binding conclusions are never applied to phage-display. The phage-display field is driven by positive results. Most publications report and follow up only on sequences enriched in the screen and consider only large copy numbers interesting. All papers focus on sequences that were found. Very few papers in phage display ask why other sequences were not found.

One of the reasons why phage display is not used for SAR-type analysis is because negative observations in phage library cannot be determined with high confidence. From a practical point of view, measuring zero with high confidence requires the largest number of observation (the highest depth of sequencing). The payoff, however, is immense: one screen with "confident zeros" could potentially yield SAR for every possible substitution of every possible amino acid. We refer to this (theoretical) possibility as "Instant SAR," and its condensed theoretical form is described in (3) or (9) and (10). This paper demonstrates that the depth of sequencing is not the only problem towards this goal. Accurate estimate of negative results requires complete characterization of the origins of errors in sequencing which yield false negative values by censoring certain sequencing. Other types of censorship, such as growth bias, should be characterized and eliminated as well. As the phage display field is currently focused on positive results, the need for optimal error corrections and recovery of erroneous reads is low. With the rise of SAR-type applications in phage display, error correction will be recognized as the most significant barrier because it could lead to improper assignment of low frequencies and negative results. Improved error correction strategies could assign a lower confidence to the sequence instead of eliminating the errors and labeling them as confident zero. Proper mathematical framework, possibly similar to the one used in this paper, could be then used to carry all confidence intervals through calculations to yield reliable SAR-type data.

We note that the framework described in this paper is suitable for the analysis of the selection from libraries in which the diversity of the libraries before and after selection could be covered entirely by deep sequencing. With the current depth of sequencing, it corresponds to medium-scale libraries of ~10^6 random members and affinity-matured libraries that contain ~10^6 point mutations. We are in the process of generating these medium-scale libraries and running selection procedures that will allow us to apply and refine our framework. In the future, as technical capabilities and depth of sequencing increase, the process would be applicable to larger libraries as well.

4. Methods

4.1. Generation of Z-Bars and Other Visualization Techniques. Sequencing of the libraries has been described in our previous publications [6, 47]. All data visualization in this paper was done by MATLAB scripts; raw *.eps output from MATLAB scripts subject to minor postprocessing in Adobe Illustrator to adjust fonts relative dimensions of plots. Core scripts are described in the supplementary information. Other scripts are available in our previous publication [47]. Illumina files used for the analysis can be found in the directory at http://www.chem.ualberta.ca/~derda/mathbiology/; the file ERROR_TAG_data0001.txt is an example of error-tagged reads; PhD7-Amp-xxF.txt is the library filtered with xx Phred cutoff (xx = 1, 13 and 30); file PhD7-Amp-13F-30F-CO.txt describes confidence intervals for Phred(13) to Phred(30) filtering process; other files with *-CO.txt extension describe confidence intervals of other processes. Supplementary Figures S1–S4 and Schemes S1 and S2 describe MATLAB implementation of the **Sa** operator.

Acknowledgments

This work was supported by funds from the University of Alberta and Alberta Glycomics Centre.

References

[1] J. K. Scott and G. P. Smith, "Searching for peptide ligands with an epitope library," *Science*, vol. 249, no. 4967, pp. 386–390, 1990.

[2] G. P. Smith and V. A. Petrenko, "Phage display," *Chemical Reviews*, vol. 97, no. 2, pp. 391–410, 1997.

[3] A. D. Ellington and J. W. Szostak, "In vitro selection of RNA molecules that bind specific ligands," *Nature*, vol. 346, no. 6287, pp. 818–822, 1990.

[4] C. Tuerk and L. Gold, "Systemic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase," *Science*, vol. 249, no. 4968, pp. 505–510, 1990.

[5] E. Dias-Neto, D. N. Nunes, R. J. Giordano et al., "Next-generation phage display: integrating and comparing available molecular tools to enable costeffective high-throughput analysis," *PLoS ONE*, vol. 4, no. 12, Article ID e8338, 2009.

[6] W. L. Matochko, K. Chu, B. Jin, S. W. Lee, G. M. Whitesides, and R. Derda, "Deep sequencing analysis of phage libraries using Illumina platform," *Methods*, vol. 58, pp. 47–55, 2012.

[7] A. Ernst, D. Gfeller, Z. Kan et al., "Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing," *Molecular BioSystems*, vol. 6, no. 10, pp. 1782–1790, 2010.

[8] G. V. Kupakuwana, J. E. Crill, M. P. McPike, and P. N. Borer, "Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing," *PLoS ONE*, vol. 6, no. 5, Article ID e19395, 2011.

[9] P. A. C. 'T Hoen, S. M. G. Jirka, B. R. Ten Broeke et al., "Phage display screening without repetitious selection rounds," *Analytical Biochemistry*, vol. 421, no. 2, pp. 622–631, 2012.

[10] H. Zhang, A. Torkamani, T. M. Jones, D. I. Ruiz, J. Pons, and R. A. Lerner, "Phenotype-information-phenotype cycle for deconvolution of combinatorial antibody libraries selected against complex systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 33, pp. 13456–13461, 2011.

[11] R. Derda, S. K. Y. Tang, S. C. Li, S. Ng, W. Matochko, and M. R. Jafari, "Diversity of phage-displayed libraries of peptides during panning and amplification," *Molecules*, vol. 16, no. 2, pp. 1776–1803, 2011.

[12] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 29, pp. 147–160, 1950.

[13] D. J. Rodi, A. S. Soares, and L. Makowski, "Quantitative assessment of peptide sequence diversity in M13 combinatorial peptide phage display libraries," *Journal of Molecular Biology*, vol. 322, no. 5, pp. 1039–1052, 2002.

[14] L. Makowski, "Quantitative analysis of peptide libraries," in *Phage Nanobiotechnology*, chapter 3, 2011.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[16] U. Ravn, F. Gueneau, L. Baerlocher et al., "By-passing in vitro screening: next generation sequencing technologies applied to antibody display and in silico candidate selection," *Nucleic Acids Research*, vol. 38, no. 21, 2010.

[17] T. Kim, M. S. Tyndel, H. Huang et al., "MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets," *Nucleic Acids Research*, vol. 40, no. 6, article e47, 2012.

[18] J. A. Weinstein, N. Jiang, R. A. White III, D. S. Fisher, and S. R. Quake, "High-throughput sequencing of the zebrafish antibody repertoire," *Science*, vol. 324, no. 5928, pp. 807–810, 2009.

[19] B. J. DeKosky, G. C. Ippolito, R. P. Deschner et al. et al., "High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire," *Nature Biotechnology*, vol. 31, pp. 166–169, 2013.

[20] B. Levitan, "Stochastic modeling and optimization of phage display," *Journal of Molecular Biology*, vol. 277, no. 4, pp. 893–916, 1998.

[21] C. S. Riesenfeld, P. D. Schloss, and J. Handelsman, "Metagenomics: genomic analysis of microbial communities," *Annual Review of Genetics*, vol. 38, pp. 525–552, 2004.

[22] M. L. Sogin, H. G. Morrison, J. A. Huber et al., "Microbial diversity in the deep sea and the underexplored 'rare biosphere'," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 32, pp. 12115–12120, 2006.

[23] F. Bäckhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson, and J. I. Gordon, "Host-bacterial mutualism in the human intestine," *Science*, vol. 307, no. 5717, pp. 1915–1920, 2005.

[24] N. Beerenwinkel and O. Zagordi, "Ultra-deep sequencing for the analysis of viral populations," *Current Opinion in Virology*, vol. 1, no. 5, pp. 413–418, 2011.

[25] J. A. Huber, D. B. Mark Welch, H. G. Morrison et al., "Microbial population structures in the deep marine biosphere," *Science*, vol. 318, no. 5847, pp. 97–100, 2007.

[26] A. Wilm, P. P. K. 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, pp. 11189–11201, 2012.

[27] P. J. Turnbaugh, C. Quince, J. J. Faith et al., "Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 16, pp. 7503–7508, 2010.

[28] C. Quince, A. Lanzén, T. P. Curtis et al., "Accurate determination of microbial diversity from 454 pyrosequencing data," *Nature Methods*, vol. 6, no. 9, pp. 639–641, 2009.

[29] S. J. Watson, M. R. A. Welkers, D. P. Depledge et al., "Viral population analysis and minority-variant detection using short read next-generation sequencing," *Philosophical Transactions of the Royal Society B*, vol. 368, no. 1614, Article ID 20120205, 2013.

[30] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow et al. et al., "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, pp. 53–59, 2008.

[31] K. Nakamura, T. Oshima, T. Morimoto et al., "Sequence-specific error profile of Illumina sequencers," *Nucleic Acids Research*, vol. 39, no. 13, 2011.

[32] M. A. Quail, I. Kozarewa, F. Smith et al., "A large genome center's improvements to the Illumina sequencing system," *Nature Methods*, vol. 5, no. 12, pp. 1005–1010, 2008.

[33] O. Zagordi, R. Klein, M. Däumer, and N. Beerenwinkel, "Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies," *Nucleic Acids Research*, vol. 38, no. 21, pp. 7400–7409, 2010.

[34] M. W. Schmitt, S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt, and L. A. Loeb, "Detection of ultra-rare mutations by next-generation sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, pp. 14508–14513, 2012.

[35] A. Syropoulos, "Mathematics of multisets," in *Proceedings of the Workshop on Multiset Processing: Multiset Processing, Mathematical, Computer Science, and Molecular Computing Points of View*, pp. 347–358, Springer, 2001.

[36] W. Arap, M. G. Kolonin, M. Trepel et al., "Steps toward mapping the human vasculature by phage display," *Nature Medicine*, vol. 8, no. 2, pp. 121–127, 2002.

[37] S. Blond-Elguindi, S. E. Cwirla, W. J. Dower et al., "Affinity panning of a library of peptides displayed on bacteriophages reveals the binding specificity of BiP," *Cell*, vol. 75, no. 4, pp. 717–728, 1993.

[38] S. Cox, E. Rosten, J. Monypenny et al., "Bayesian localization microscopy reveals nanoscale podosome dynamics," *Nature Methods*, vol. 9, no. 2, pp. 195–200, 2012.

[39] E. Rosten, G. E. Jones, and S. Cox, "Image plug-in for Bayesian analysis of blinking and bleaching," *Nature Methods*, vol. 10, pp. 97–98, 2013.

[40] W. L. Matochko, S. Ng, M. R. Jafari, J. Romaniuk, S. K. Y. Tang, and R. Derda, "Uniform amplification of phage display libraries in monodisperse emulsions," *Methods*, vol. 58, pp. 18–27, 2012.

[41] E. A. Peters, P. J. Schatz, S. S. Johnson, and W. J. Dower, "Membrane insertion defects caused by positive charges in the early mature region of protein pIII of filamentous phage fd can be corrected by prlA suppressors," *Journal of Bacteriology*, vol. 176, no. 14, pp. 4296–4305, 1994.

[42] L. A. Brammer, B. Bolduc, J. L. Kass, K. M. Felice, C. J. Noren, and M. F. Hall, "A target-unrelated peptide in an M13 phage display library traced to an advantageous mutation in the gene II ribosome-binding site," *Analytical Biochemistry*, vol. 373, no. 1, pp. 88–98, 2008.

[43] G. A. Kuzmicheva, P. K. Jayanna, I. B. Sorokulova, and V. A. Petrenko, "Diversity and censoring of landscape phage libraries," *Protein Engineering, Design and Selection*, vol. 22, no. 1, pp. 9–18, 2009.

[44] D. R. Wilson and B. B. Finlay, "Phage display: applications, innovations, and issues in phage and host biology," *Canadian Journal of Microbiology*, vol. 44, no. 4, pp. 313–329, 1998.

[45] R. Derda, S. K. Y. Tang, and G. M. Whitesides, "Uniform amplification of phage with different growth characteristics in individual compartments consisting of monodisperse droplets," *Angewandte Chemie*, vol. 49, no. 31, pp. 5301–5304, 2010.

[46] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing," *Nucleic Acids Research*, vol. 36, no. 16, article e105, 2008.

[47] W. L. Matochko, S. C. Li, S. K. Y. Tang, and R. Derda, "Prospective identification of parasitic sequences in phage display screens," *Nucleic Acids Research*, 2013.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 943636, 11 pages
http://dx.doi.org/10.1155/2013/943636

Review Article

Bioinformatics Resources and Tools for Conformational B-Cell Epitope Prediction

Pingping Sun,[1,2] Haixu Ju,[1,3] Zhenbang Liu,[1,3] Qiao Ning,[1] Jian Zhang,[1,3] Xiaowei Zhao,[1,3] Yanxin Huang,[2] Zhiqiang Ma,[1,3] and Yuxin Li[2]

[1] *School of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China*
[2] *National Engineering Laboratory for Druggable Gene and Protein Screening, Northeast Normal University, Changchun 130024, China*
[3] *Key Laboratory of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun 130117, China*

Correspondence should be addressed to Zhiqiang Ma; mazq@nenu.edu.cn and Yuxin Li; liyx486@nenu.edu.cn

Received 19 March 2013; Revised 22 May 2013; Accepted 1 June 2013

Academic Editor: Jian Huang

Identification of epitopes which invoke strong humoral responses is an essential issue in the field of immunology. Localizing epitopes by experimental methods is expensive in terms of time, cost, and effort; therefore, computational methods feature for its low cost and high speed was employed to predict B-cell epitopes. In this paper, we review the recent advance of bioinformatics resources and tools in conformational B-cell epitope prediction, including databases, algorithms, web servers, and their applications in solving problems in related areas. To stimulate the development of better tools, some promising directions are also extensively discussed.

1. Introduction

A B-cell epitope is defined as a region of an antigen recognized by either a particular B-cell receptor (BCR) or subsequently the elicited antibody in a humoral response [1–3]. A B-cell epitope can be categorized into two types by its spatial structure: liner epitope or conformational epitope. A liner epitope (also called continuous epitopes) is composed of residues that are sequentially consecutive, whereas a conformational epitope (also known as discontinuous epitope) consists of sequential segments that are brought together in spatial proximity when the corresponding antigen is folded. It has been reported that more than 90% of B-cell epitopes are discontinuous B-cell epitopes [4, 5].

The identification of B-cell epitopes is rather important to immunodetection and immunotherapeutic applications since an epitope as the minimal immune unit is strong enough to elicit a potent humoral immune response with no harmful side effects to human body [3, 6]. The ultimate goal of epitope prediction is to aid the design of molecules that can mimic the structure and function of a genuine epitope and replace it in medical diagnostics and therapeutics and also in vaccine design [2, 7]. The most reliable methods for identification of an epitope are X-ray crystallography and NMR techniques [8, 9], but they are time consuming and expensive. Hence, computational methods and tools, with the virtues of low cost and high speed, were employed to predict B-cell epitopes in silico.

The interaction between an antigen and an antibody is a complicated biochemical process. An antibody, which has a "Y"-shape structure, binds to the epitopic region of an antigen through a highly variable complementarily determining region (CDR). The interaction between an antigen and an antibody is mainly through the connections of intermolecular low energy (e.g., hydrogen bond, hydrophobic interaction, and van der Waals force) and few connections of intermolecular high energy (e.g., salt bridge). Moreover, since an antibody interacts with an antigen through a deep and narrow antigen-binding clef, it is reasonable to believe that the interaction between an antigen and an antibody involves both specific sequence recognition and mutual structure identification.

By far, the study of B-cell epitope prediction mainly aimed at predicting linear epitopes [10–24]. However, since most B-cell epitopes are conformational epitopes, the prediction of liner B-cell epitope has limited application. In recent years,

Table 1: Databases for 3D structure of the antigen and epitopes data.

Databases	Websites
PDB [31]	http://www.rcsb.org/pdb/home/home.do
CED [32]	http://immunet.cn/ced/
IEDB [33]	http://www.immuneepitope.org/
HIV Molecular Immunology [35]	http://www.hiv.lanl.gov/content/immunology/index

some computational methods were proposed though the number is limited and the performance is not significant [25–29]. Consequently, to improve the performance of B-cell epitope prediction, integrating multidisciplinary knowledge and combining different methods become a promising prospective.

In this work, we review recent advances in computational methods for conformational B-cell epitopes prediction, including databases, algorithms, web servers, and their applications, point out some problems in the current state of the art, and outline some promising directions for improving the prediction of conformational B-cell epitopes.

2. Structure-Based Prediction Methods

B-cell epitopes prediction based on the 3D structure of antigen began in 1999 [30], and the core idea of the prediction methods is through the 3D structure of antigen and epitope-related propensity scales, including geometric attributes and specific physicochemical properties. In recent years, with the development of various omics and bioinformatics, related experimental data of conformational B-cell epitopes has been accumulating rapidly. The development of epitope-related databases promotes conformational B-cell epitopes prediction. Herein, we review the major databases and approaches for predicting conformational B-cell epitopes based on the 3D structure of an antigen.

2.1. Databases. The availability of experimental data plays a pivotal role in conformational B-cell epitope prediction. The 3D structure of antigen or the complex of antigen-antibody is stored in the PDB database [31], and the data for epitopes and other associate information were stored in some special databases. Table 1 lists all the epitope-related databases together with their functional comments.

PDB [31] database compiles the compounds derived from the X-ray crystallography and NMR experiments. The majority of the information from PDB database are the 3D structure of protein. One can search needed structure according the PDB-id in the home page and then view or download the structure in several formats. CED database [32] comprising the annotated epitopes which was determined by experimental methods. The database provides a user-friendly web interface, and most epitopes in database can be viewed interactively in the context of their 3D structures. One can browse all the entries or search the certain entry from the corresponding hyperlinks in the home page. IEDB database is the most commonly used and most authoritative database in epitope prediction [33, 34]. Since IEDB 2.0 released, there were 38,552 entries on B-cell epitope and a handful of integrated prediction tools providing much convenience for researchers. Researchers can search interested B-cell epitope from the pull-down menu of "Advanced search" on the home page. HIV Molecular Immunology database contains HIV virus epitopes which were determined by experiments [35]. Both the B-cell epitopes and T-cell epitopes are included. This database provides convenience for the research of specific HIV virus epitopes.

The previous databases are important resources for conformational B-cell epitope prediction. The data from these databases provide a basis for computational biologists to derive benchmark and customize datasets for new algorithm development and tool evaluation.

2.2. Algorithms, Programs, and Their Application. Comparing with mimotope-based prediction methods which will be introduced in what follows, structure-based methods for conformational B-cell epitopes prediction have the advantage that they only need the structure of antigen. In 1999, Kolaskar and Kulkarni-Kale used the 3D structure of antigen to analyse and locate the conformational epitopes of Japanese encephalitis virus by calculating the surface accessible fragments of amino acids [30]. They improved the algorithm and released CEP which is the first web-based software for conformational epitope prediction in 2005 [36]. The essential ideal of CEP is to generate surface fragments of an antigen, and then use the spatial distance of these fragments and other statistical characteristics to locate epitopes. The structure-based algorithms, web servers (programs) and brief notes are listed in Table 2.

DiscoTope was the second web-based conformational epitope prediction software [37]. In 2006, Andersen et al. collected a dataset which contains 76 antigen-antibody complexes. To investigate the role of certain features that distinguish epitopes from nonepitopes, a number of statistics were studied including the distribution of length and segments of an epitope and single amino acid preference and Parker hydrophilicity. Through a combination of statistics, spatial context, DiscoTope could successfully predict the location of epitopes on the previously mentioned dataset. In 2007, Rapberger et al. proposed a new kind of conformational B-cell epitopes prediction framework [38]. They took advantage of the complementary geometric shape of antigen epitopes and antibody paratope, as well as the measure of binding energy of antigen and antibody. The method was the first one which considered the antibody information in the research of epitopes prediction.

The first conference for B-cell epitope prediction was held in Washington 2007. The meeting published a benchmark dataset for conformational B-cell epitope prediction in

Table 2: Methods for structure-based B-cell epitopes prediction.

Method	Online service (program)	Brief notes (features used in prediction method)
CEP [30, 36]	Available upon request	First Web-based conformational B-cell epitopes prediction software, based on the surface accessible
DiscoTope [37]	http://www.cbs.dtu.dk/services/DiscoTope/	Amino acid statistics, spatial context, and surface accessibility
Rapberger[a] [38]	Not stated	Based on antibody information
ElliPro [40]	http://tools.immuneepitope.org/tools/ElliPro/iedb_input	Prominent index
PEPITO/BEPro [41]	http://pepito.proteomics.ics.uci.edu/	Half sphere exposure values
PEPOP [42]	Available upon request	Accessible and sequence contiguous amino acids segments
SEPPA [45]	http://lifecenter.sgst.cn/seppa/index.php	Unit patch of residue triangle
Epitopia [46]	http://epitopia.tau.ac.il/	Based on Naïve Bayes classifier with physicochemical and structural geometrical properties
EPCES [47]	http://sysbio.unl.edu/services/EPCES/	Consensus score by six functions
Shinji Soga[a] [49]	Not stated	Antibody-specific epitope propensity index
EPSVR & EPMeta [51]	http://sysbio.unl.edu/services/	Based on SVR and meta-analysis
Zhang[a] [52]	http://code.google.com/p/my-project-bpredictor/downloads/list	Based on random forests with a distance-based feature

[a]The name of the first author is used if the method has no name.

the format of the 3D structure of antigen chosen from PDB database. The benchmark dataset includes 62 3D structures of antigens with inferred epitopes. The construction of this benchmark dataset accelerated the development of conformational B-cell epitopes prediction and provided a basis for method evaluation. Ponomarenko and Bourne evaluated CEP and DiscoTope using the benchmark dataset in the same year [39]. The results indicated that the performance of both methods did not exceed 40% of precision and 46% of recall. Consequently, methods with better performance are still in great need. One way to attain this goal is through developing new features and combining them.

In the next few years, newly proposed conformational B-cell epitope prediction methods managed to look for effective propensity scales or combine the available amino acid physicochemical properties and geometrical structure properties. In 2008, three conformational B-cell epitope prediction methods were proposed: ElliPro [40], PEPITO [41], and PEPOP [42]. The main idea of ElliPro attributes to the liner B-cell epitopes prediction method of Thornton et al. [43]. ElliPro predicts conformational B-cell epitopes by combining the geometric features of an antigen and single amino acid epitope propensity. When the structure is not available, ElliPro first model the 3D structure of the antigen by searching for its homologues in PDB or running MODELLER [44]. PEPITO predicts conformational B-cell epitopes using a combination of single amino acid epitope propensity and half sphere exposure values at multiple distances. One major improvement of PEPITO is that it employed half sphere exposure to describe the degree of compactness which inspired the latter methods. PEPOP identifies segments composed of accessible and sequentially contiguous amino acids of the 3D structure of an antigen and then clusters these segments according to their spatial distances to identify epitopes. Another contribution of PEPOP is designing immunogenic peptides through the results of epitopes identification.

SEPPA [45], Epitopia [46], and EPCES [47] were published in 2009. SEPPA employs the concept of "unit patch of residue triangle" to describe the local spatial context of protein surface and "clustering coefficient" to describe the spatial compactness of surface residues. Then, the two features are combined to predict epitopes. Epitopia adopts the idea of partition which divides a given antigen to overlapping surface patches. Then, the scores of physicochemical and structural-geometrical properties for central residue of each patch are calculated before using a Naïve Bayes classifier to predict the immunogenic potential of protein regions. EPCES proposed six epitopes propensities, including conservation score, side-chain energy score, contact number, surface planarity score, and secondary structure composition. With the vote mechanism, EPCES reaches a consensus score which represents the likelihood of being an epitope based on the scale of each feature. Based on the features, we trained an SVM classifier to predict conformational epitopes [48], and the testing results showed that different classification methods did not improve the accuracy of the prediction performance based on these propensities.

To develop better features, Soga et al. emphasized information hidden in antibody in the process of antigen and antibody interactions [49]. They defined the antibody-specific epitope propensity (ASEP) index. Then, it was used to predict

epitopes together with the result from DiscoTope. This paper made the first attempt to identify epitopes by combining different prediction methods. In 2011, Sun et al. collected a latest comprehensive dataset and did detailed statistical analysis of epitope residues and nonepitope residues from several aspects [50]. The study of antigen and antibody interaction pattern revealed the importance of antibody information in epitopes prediction as well. In the same year, two novel server applications EPSVR and EPMeta were presented by the same author of EPCES [51]. EPSVR uses a support vector regression method to integrate six scoring terms as EPCES, while EPMeta is a metaserver which combined with EPSVR, EPCES, Epitopia, SEPPA, PEPITO, and Discotope1.2. In 2011, Zhang et al. proposed a new epitope prediction method [52]. The method proposed a concept of "thick surface patch" which brought the impact of interior residues, the adjacent residue distance feature, into consideration. It reflects the unequal contributions of adjacent residues to the location of binding sites and the random forest algorithm which is used to process imbalanced data. The method represented higher prediction accuracy comparing with other methods.

The structure-based conformational B-cell epitopes prediction methods are all based on the structure features of antigen, and a different method employs different propensity scales. Most of the previously mentioned prediction methods offer online service or program (see Table 2). The online services have a user-friendly interface. The usage of these methods is simple. Researchers enter the PDB ID or upload the local file in PDB format, determine the antigen chain, and specify the corresponding thresholds according to the orders that will later get the prediction results. Yao et al. [53] construct a benchmark and evaluate the performance of all existing prediction methods. The results show that the accuracy of EPMeta is the overall highest value by all conditions and methods. It states that in the case of different prediction methods usually not give a consensus result, and consider the results of the multiple prediction methods is a better choice.

2.3. Current Problems. B-cell epitope prediction based on the 3D structure of antigen structure has already made some progress, even so the methods need further improvements. Firstly the dataset, which is essential for the methods based on machine learning, is relatively small and inconsistent. Moreover, since nonepitopic amino acids are defined as the amino acids which are not a part of currently determined epitopes, the undetermined epitopic amino acids would very likely bring in noises in the process of statistical learning. In addition, the input and output formats for each method is different which make it difficult to evaluate the performance of different methods.

Secondly, in order to assess the validity and performance of the prediction methods, both antigen structure and the epitope information are needed. CED and IEDB annotated epitope sites for part of structures, and we call this annotated epitopes which are actually determined by wet experiment as functional epitopes. But this situation is not the same for the other structures. To use these structures, one needs to determine the epitope of the structures by distance between antigen and antibody or accessible surface area ((ASA), and Surface Racer [54] and NACCESS [55] are commonly used tools that are designed for calculating ASA) loss upon antibody binding at first, and we call this kind of epitope as structure epitopes. The difference in epitopes determination makes prediction methods producing relatively poorer performance on the structure epitopes-based datasets than on the functional epitopes based datasets.

Lastly, an antibody binds to an antigen by the spatial structure, so there is a wealth of information hidden in the 3D structure of antigen and antibody. Theoretically, the features extracted from the structure would certainly improve the performance of existing B-cell epitopes prediction methods. However, it is more complicated to extract features from the 3D structure of an antigen than dealing with the primary sequence. Features mentioned in these papers do not have enough ability to distinguish the epitopic residues from the rest.

3. Mimotope-Based Prediction Methods

Mimotope-based prediction is a combinatorial method which requires both antibody affinity-selected peptides and the 3D structure of antigen as input. To attain affinity-selected peptides, random peptides are initially displayed on the surface of filamentous phages. Then, random peptides which bind to a monoclonal antibody with a certain degree of affinity are screened, eluted, and amplified. After 3–5 rounds of the operation, the resulting peptides become fewer but with higher affinity. These affinity-selected peptides are defined as mimotopes. Mimotopes and genuine epitopes can combine the same paratope of monoclonal antibody and cause immune response, so they have the similar functionality with the genuine epitope [56, 57]. Besides, the selected mimotopes commonly share high sequential similarity which implies that certain key binding motifs and physicochemical preferences exist during the interaction. Therefore, mapping these mimotopes back to the source antigen can help finding the genuine epitopes more accurately. In what follows, we review the major databases and approaches for predicting conformational B-cell epitopes based on mimotopes.

3.1. Databases. Mimotope-based methods need both the structure of antigen and the sequence data of mimotopes. Since the 3D structure of the antigen can be obtained from PDB or by computational homology modeling, the small number of mimotope sequences derived from phage display becomes a limitation for the development of conformational B-cell epitopes prediction based on mimotopes. In recent years, several databases which integrated the structure data, the mimotopes data, and other associate information have been released which play a fundamental role in Immunoinformatics. Table 3 lists current databases which contain the information of mimotope.

ASDP was a curated database that incorporated data on full-length protein, proteins, protein domains, and peptides which were obtained mainly from phage display experiment [58]. It was the first database for mimotopes. The current

Table 3: Available databases for mimotopes data.

Databases	Websites
ASPD [58]	http://wwwmgs.bionet.nsc.ru/mgs/gnw/aspd
RELIC Peptides [59]	Available upon request
PEPBANK [60]	http://pepbank.mgh.harvard.edu
MimoDB [61, 62]	http://immunet.cn/mimodb
Sun's Benchmark datasets [25]	http://cs.nenu.edu.cn/bioinfo/benchmark%20datasets/index.html

version released in 2001 has 195 entries. ASPD has a user-friendly interface, and researchers can search the needed information by means of the SRS system. The RELIC Peptides is a relational database that contains more than 5,000 peptide sequences selected with small molecule metabolites drugs as well as random clones from parent libraries [59]. RELIC Peptides is indispensable as part of the RELIC suite for many tools in RELIC depend on the data. PepBank is a database of peptides based on sequential text mining and public peptide data sources [60]. This database stores peptides with available sequences and the length equals 20 amino acids or shorter. PepBank has a web-based user interface with a simple, Google-like search function, advanced text search, and BLAST and Smith-Waterman search capabilities. MimoDB is an information portal to biopanning results of random libraries [61, 62]. It is the latest and largest database for mimotopes. In version 2.0, it has 15,633 peptides collected from 849 papers and groups into 1,818 sets. For each entry, the target, template, library, and structures information are given. In addition, MimoDB provides tools for simple and advanced search, structure visualization, BLAST, and alignment view on the fly.

Sun's benchmark datasets were constructed by our team in 2011, and it is special for conformational B-cell epitope prediction based on mimotope analysis. Now, we have established benchmark 2.0 already. The benchmark 2.0 consists of 39 complex structures with 66 mimotope sets; the 39 complex structures contain 16 antigen-antibody complexes and 23 protein-protein interactions structures. In addition, we provide 24 test cases as representative datasets which have only one mimotope set for one complex structure. Each set includes the complex structure, the template chain, the mimotopes obtained from corresponding phage display experiment, and the epitope information. All the datasets can be downloaded freely for academic purposes. The benchmark dataset can be freely accessed at http://cs.nenu.edu.cn/bioinfo/benchmark%20datasets/index.html.

The databases described previously are important resources for the mimotope-based B-cell epitope prediction. With the large amount of mimotopes in these databases as well as the protein structure databases, it is feasible to construct a benchmark for development and evaluation of new mimotope-based epitope prediction methods.

3.2. Algorithms, Programs, and Their Application. Mimotope-based prediction methods are essential to map mimotopes back to the surface of a source antigen to locate the best alignment sequences and predict possible epitopic regions. The available mimotope-based algorithms, web servers (programs), and brief notes are listed in Table 4.

Huang et al. classified the mimotope-based epitope prediction into two categories: one is the sequence-sequence alignment methods and the other is sequence-structure alignment methods [63]. Among the prediction methods listed previously, FINDMAP, EPIMAP, and the MimAlign algorithm of MIMOP belong to the sequence-sequence alignment methods. The inputs of these methods are mimotopes and the primary structure of an antigen. FINDMAP aligns the motif extracted from mimotopes to the antigen sequence directly rating the best matching sequences as epitope candidates [64]. EPIMAP is an improved version of FINDMAP [65]. It aligns each mimotope to the antigen sequence and then selects the most mutually compatible alignments from a set of the top-scoring alignments before filtering out spurious alignments with EPIFILTER program. MIMOP was proposed by Moreau et al. in 2006 [66] which includes two parts: MimAlign and MimCons. MimAlign combines results from four multiple sequence alignments of the antigen and mimotopes sequences in a combined alignment. For each position of the combined alignment, a frequency and a score are calculated. Convergent positions are then selected and clustered based on their topology. The clusters attained are considered as potential epitopic regions.

The remaining methods belong to the sequence-structure alignment methods. Further, Huang classified these methods into 5 kinds according to the mean of sequence-structure alignment [63]: motif-based methods, pairs-based methods, patch-based methods, graph-based methods, and hybrid methods.

The motif-based methods aim to obtain motif through multiple alignment of mimotopes and then map the motif to the surface of an antigen to locate B-cell epitopes. MEPS, 3DEX, MIMOX, and the MimCons algorithm of MIMOP belong to this kind. MEPS is the first B-cell epitope predicting method based on mimotope analysis [67]. MEPS first model an antigen surface into fixed-length peptides and then aligns each of the short peptide to the motif derived from multiple alignment of the mimotopes. The best aligned short peptides are treated as candidate epitopes. 3DEX takes the physicochemical neighborhood of Cα- or Cβ-atoms of individual amino acids into account [68]. A given amino acid in a peptide sequence is localized by the protein, and the software searches within predefined distances for the amino acids neighboring that amino acid in the peptide. Surface exposure of amino acids can also be taken into consideration.

Table 4: Methods for mimotope-based B-cell epitopes prediction.

Method	Online service (program)	Brief notes
FINDMAP [64]	Not available	Map motif to antigen sequence
EPIMAP [65]	Not available	Improved version of FINDMAP
MIMOP/MimAlign [66]	Available upon request	Based on four multiple sequence alignments
MEPS [67]	http://www.caspur.it/meps	Surface mimicking peptides
3DEX [68]	Not available	Physicochemical neighborhood
MIMOX [88]	http://immunet.cn/mimox/	The first free web tool
MIMOP/MimCons [66]	Available upon request	Clustering the mimotope sequences
Mapitope [69, 70]	http://pepitope.tau.ac.il	The first method based on amino acid pairs
Denisova[a] [71–73]	Not available	Derivative method of Mapitope
SiteLight [74]	Not available	A patch-based method
EpiSearch [75]	http://curie.utmb.edu/episearch.html	An automated sequence analysis based on sequence and 3D profiles
PepSurf [76]	http://pepitope.tau.ac.il	The first graph-based method
Pep-3D-Search [77]	http://kyc.nenu.edu.cn/Pep3DSearch	Ant colony optimization algorithm
MimoPro [78]	http://informatics.nenu.edu.cn/MimoPro	Based on both patch and graph searching

[a]The name of the first author is used if the method has no name.

The procedure is then repeated for the remaining amino acids of the peptide. This procedure may cost few hours. MIMOX is the first freely accessible web tool for mimotope-based B-cell epitope prediction [62]. It has two parts. The first part provides a simple interface for the alignment of mimotope sets, while the second part of MIMOX maps a single mimotope or a motif derived from the first part onto the corresponding antigen and rates all of the clusters of residues to locate the genius epitope. MimCons is another part of MIMOP method, and it evaluates the similarity of the mimotope sequences and clusters them accordingly. Motifs are identified from mimotope sequences of each cluster. The accessible surface of the antigen is scanned to find out all possible exposed consensus patterns. Spatial neighbor amino acids are identified and constitute potential epitopes. In addition, MimAlign and MimCons can be run either independently or with their results combined.

The essential idea of pairs-based methods is to predict B-cell epitopes with the statistical characteristics of amino acid pairs. Mapitope and Denisova belong to this kind. In 2003, Enshell-Seijffers et al. described a mimotope-based approach to predict the epitopes of the HIV-1 [69]. Firstly, they defined amino acid pairs (AAP) with a predefined distance threshold between the central carbon atom of two neighbor residues. Secondly, they defined statistically significant pairs (SSPs) by calculating the probabilities of each AAP. Lastly, the SSPs are mapped to the 3D structure of an HIV-1 antigen to locate epitopes. In 2007, Bublil et al. applied this method to conformational B-cell epitope prediction and presented the tool as Mapitope [70]. A continuous work by Denisova et al. took all possible space pairs, including pairs separated by one residue, two residues, three residues, and so on in mimotopes into account and identified epitopes by pattern recognition theory [71–73]. This method is specially designed for elucidating epitope specificity within antiserum.

The core idea of patch-based methods is dividing the surface of antigen into overlapping patches and selecting high-scored amino acid residues as candidate epitopes by comparing mimotopes with patches based on sequence similarity or the statistical characteristics of amino acids. SiteLight and EpiSearch belong to this category. SiteLight divides the antigen surface into overlapping patches, and then aligns each mimotope to each of the patches. To identify candidate epitopes, the best matched paths are selected repeatedly until 25% of antigen surface is covered [74]. EpiSearch predicts conformational B-cell epitopes by an automated sequence analysis of mimotopes and a comparison to the distribution of amino acids on patches on the antigen surface [75]. The amino acid compositions of the mimotopes and 3D profile of an antigen are compared and quantified in a score function for each patch on the antigen surface. The highest scoring patches are listed in the output files and are also displayed on the surface of the protein.

The main idea of graph-based methods is to model the amino acids from an antigen as a graph structure so as to use the graph search methods to locate potential epitopes. PepSurf and Pep-3D-Search belong to this category. PepSurf searches the best matched paths from the graph built from the antigen with mimotope sequences using color-coding algorithm and dynamic programming algorithm [76]. Pep-3D-Search searched for the matched paths on the antigen surface by the Ant Colony Optimization (ACO) algorithm [77]. Candidate epitopes were then formed by clustering the resulting paths with a high P value score by the Depth-First Search algorithm. Pep-3D-Search provides two modes of B-cell epitope prediction: (1) mimotope-based search and (2) motif-based search.

The last kind of mimotope-structure alignment B-cell epitope prediction method is a hybrid method. MimoPro, which was proposed by our team in 2011, is the first attempt to integrate the idea of different methods. The method employs

the idea of patch-based and graph-based searching [78]. The core of MimoPro is a searching algorithm operated on a series of overlapping patches on the surface of antigen. These patches are then transformed to a number of graphs using an adaptable distance threshold (ADT) regulated by compactness factor (CF), a novel parameter proposed in this method. Then on each single patch, a complete search is conducted to guarantee the best alignment for each mimotope sequence. Dynamic programming and branch-bound methods are also adopted to both avoid repetition in searching and further narrow the search space.

Unfortunately, the available service of the previous 14 methods is few. At present, there are only three available freely web-based B-cell epitope prediction service platforms in the world. The first is PEPITOPE [79], and it provides online service based on three methods: Mapitope, PepSurf, and the combined. The web service of the three methods has the restriction that the length of mimotope sequence cannot be longer than 14 amino acids. Besides, Mapitope and PepSurf also can be run in local, and the local version has no service restriction. The second is EpiSearch and the epitope prediction method is EpiSearch only [75]. EpiSearch has the restriction that the number of mimotope sequences cannot exceed 30 amino acids. The third prediction platform is PepMapper which is released by our team in May 2012 [80]. PepMapper also provides online service based on three methods: Pep-3D-Search, MimoPro, and the combined. Since Pep-3D-Search is based on the establishment of empirical background distribution for aligning score of every mimotope and antigen, and if the P value of aligning score for every mimotope is bigger than 10^{-3}, Pep-3D-Search will not give any prediction result. Among all these methods, only MimoPro has no limitation. As the structure-based conformational B-cell epitopes prediction methods, a different method employs different prediction strategy, and will not give a completely consensus prediction result. As Liang's idea [53], we think meta-analysis may be a better solution, and we are engaged in certifying this idea now.

3.3. Current Problems. Mimotope-based B-cell epitope prediction methods located epitopic region through the information from mimotopes which is obtained from experimental methods. Mimotope-based prediction is statistically more accurate, but it requires the information of mimotopes from experimental data. However, comparing with X-ray crystallography and NMR methods, in vitro screening methods have a low price to pay. Moreover, the methods can locate the interacting epitope in a designate antigen-antibody interaction context.

Despite that, accurate prediction of epitopes is still a long way to go. In 2011, we constructed a benchmark dataset for conformational B-cell epitope prediction and evaluated five mimotope-based prediction software products [25]. The result showed that in no method did the performance exceed a 0.42 of precision and 0.37 of sensitivity. The poor performance of the prediction is rooted in several aspects. The size and diversity of the benchmark dataset is inadequate, as well as many problems in mimotope-based B-cell epitope prediction need to be further studied. MimoPro combines the idea of different methods. By employing a novel idea of ADT which reflects the flexibility of interaction between amino acid pairs, MimoPro reached is the highest sensitivity among the methods, but the overall performance is still not satisfactory. How to express conformational changes in the interactions of antigen and antibody, how to establish rational mathematical model through integration of mimotopes information and the statistical characteristics of amino acids, and design intelligent search algorithm on the surface of antigen are the main directions to further improve the performance of mimotope-based B-cell epitope prediction methods.

4. Other Methods

In this section, we will focus on the development of other conformational epitope prediction methods aside from the structure-based methods and the mimotope-based methods.

4.1. Sequence-Based Methods. Sequence-based prediction methods only rely on the primary sequence of an antigen and inherit the idea of liner B-cell epitopes prediction. Particularly, the methods employs propensity scales to measure the probability of each residue being part of epitopes [37]. To reduce fluctuations, sliding window strategy is usually used.

In 2010, Ansari and Raghava proposed a method to predict conformational B-cell epitopes from the primary sequence of antigen [81]. In the method, sparse encoding scheme (BPP), physicochemical features (PPP), and amino acid composition (CCP) are extracted from the overlapping amino acid segments sliced from antigen sequences and used to train a SVM for prediction. There are two newly published methods that predict conformational B-cell epitopes by antigen sequence in last year. The two methods are BEST [82] and Zhang's [83] method. They all extract enough sequence characters first, and then BEST method employed SVM for classification, while Zhang's method adopted the ensemble learning approach to handle various features for epitope prediction.

As the high experimental requirements for resolution of protein 3D structure, the 3D structure of a large number of protein has not been resolved, and the B-cell epitope prediction methods based on antigen sequence may be worth more deeper research. Compared with structure-based prediction methods, the performance of sequence-based methods did not improve a lot, but the thought of sequence-based methods provides innovative research ideas for conformational B-cell epitope prediction.

4.2. Binding Sites Prediction Methods. The interaction of an antigen and an antibody is a subtype of protein-protein interaction, so some methods that focus on binding sites prediction of protein-protein interaction can be borrowed for conformational B-cell epitopes prediction. Recently, Yao et al. [53] construct a benchmark and evaluate the performance of all existing structure-based B-cell prediction methods, along with 4 binding sites prediction methods: ProMate [84], ConSurf [85], PINUP [86], and PIER [87]. The results

showed that the performances of the binding site prediction methods to predict B-cell epitopes are significantly lower than all structure-based epitope prediction methods. In fact, the interaction between antigen and antibody is different from other kinds of protein-protein interaction in some degree. For instance, protein-protein binding sites are usually more conserved than other surface residues to maintain the functionality of the protein, while the antigen-antibody binding sites (epitope) are less conserved due to the competition for survival against the host immune system. Hence, using these prediction methods for epitopes prediction has certain drawbacks. More importantly, the prediction methods need both the antigen and antibody structure, but epitope prediction methods are designed to identify the potential epitopes on the antigen when the antibodies are unknown. However, the epitope prediction of unbound structure has more practical value in general. Due to the different purposes, the binding sites prediction methods have little advantage in epitope prediction.

5. Conclusions and Prospects

B-cell epitope prediction is important for vaccine design, development of diagnostic reagents, and interpretation of the antigen-antibody interactions on a molecular level. In recent years, with the development of various omics and bioinformatics, related experimental data of conformational B-cell epitopes has been proposed rapidly. The construction of relevant databases promote the development of conformational B-cell epitopes prediction. In this study, we make a systematic review about the bioinformatics resources and tools for conformational B-cell epitope prediction. Though the developments, the overall performance is still not satisfactory. In what follows, we point out several aspects that may improve the performance of conformational B-cell epitopes prediction.

Build Large and Reliable Datasets. Areliable dataset should meet the requirement of nonredundant antigen structures (bound or unbound), well-defined B-cell epitopes, and the mimotope sequences. Nonredundant and abundant datasets could avoid the performance of B-cell epitope prediction methods overly optimistic. Well-defined B-cell epitopes is the premise of epitope relevant feature extraction and directly impacts the prediction performance. Mimotopes sequence is especially important for the mimotope-based conformational B-cell epitope prediction. Furthermore, large and reliable datasets are important for both training and testing. Training datasets are used to feature extraction and model training, while testing datasets is responsible for testing the performance of prediction method and evaluating the performance between different methods.

Extracting Effective Epitope Relevant Features. The essence of structure-based conformational B-cell epitope prediction is pattern classification. Extracting effective epitope relevant features is the most important part in structure-based conformational B-cell epitope prediction methods which is also the key point in B-cell epitope predictions. By far, there is no single feature or combination of features that can effectively distinguish epitopes from nonepitopes. To improve the performance of conformational B-cell epitope prediction methods, selecting effective features, or feature combination as well as integrating the mimotope-based methods may be a promising area.

Devise Intelligent Searching Algorithms. The essence of mimotope-based conformational B-cell epitope prediction is searching similar sequences with mimotopes on the surface of antigen. Intelligent searching algorithms could improve the effectiveness of the methods, as well as the prediction performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61172183), the Science Foundation for Young Teachers of Northeast Normal University (no. 12QNJJ005), the 2012 postdoctoral research projects of Jilin province, and the Scientific and Technical Project of Administration of Traditional Chinese Medicine of Jilin province (2011-zol16).

References

[1] M. H. van Regenmortel, "The concept and operational definition of protein epitopes," *Philosophical transactions of the Royal Society of London B*, vol. 323, no. 1217, pp. 451–466, 1989.

[2] 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.

[3] O. M. R. Westwood and F. C. Hay, *Epitope Mapping: A Practical Approach*, Oxford University Press, Oxford, UK, 2001.

[4] M. H. V. van Regenmortel, "Antigenicity and immunogenicity of synthetic peptides," *Biologicals*, vol. 29, no. 3-4, pp. 209–213, 2001.

[5] D. J. Barlow, M. S. Edwards, and J. M. Thornton, "Continuous and discontinuous protein antigenic determinants," *Nature*, vol. 322, no. 6081, pp. 747–748, 1986.

[6] M. B. Irving, O. Pan, and J. K. Scott, "Random-peptide libraries and antigen-fragment libraries for epitope mapping and the development of vaccines and diagnostics," *Current Opinion in Chemical Biology*, vol. 5, no. 3, pp. 314–324, 2001.

[7] M. J. Gómara and I. Haro, "Synthetic peptides for the immunodiagnosis of human diseases," *Current Medicinal Chemistry*, vol. 14, no. 5, pp. 531–546, 2007.

[8] J. J. Rux and R. M. Burnett, "Type-specific epitope locations revealed by X-ray crystallographic study of adenovirus type 5 hexon," *Molecular Therapy*, vol. 1, no. 1, pp. 18–30, 2000.

[9] M. Mayer and B. Meyer, "Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor," *Journal of the American Chemical Society*, vol. 123, no. 25, pp. 6108–6117, 2001.

[10] M. Levitt, "A simplified representation of protein conformations for rapid stimulation of protein folding," *Journal of Molecular Biology*, vol. 104, no. 1, pp. 59–107, 1976.

[11] T. P. Hopp and K. R. Woods, "Prediction of protein antigenic determinants from amino acid sequences," *Proceedings of the*

National Academy of Sciences of the United States of America, vol. 78, no. 6 I, pp. 3824–3828, 1981.

[12] J. M. R. Parker, D. Guo, and R. S. Hodges, "New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites," *Biochemistry*, vol. 25, no. 19, pp. 5425–5432, 1986.

[13] E. Westhof, D. Altschuh, and D. Moras, "Correlation between segmental mobility and the location of antigenic determinants in proteins," *Nature*, vol. 311, no. 5982, pp. 123–126, 1984.

[14] J. L. Pellequer, E. Westhof, and M. H. V. van Regenmortel, "Predicting location of continuous epitopes in proteins from their primary structures," *Methods in Enzymology*, vol. 203, pp. 176–201, 1991.

[15] J. Janin, "Surface and inside volumes in globular proteins," *Nature*, vol. 277, no. 5696, pp. 491–492, 1979.

[16] G. W. Welling, W. J. Weijer, R. van der Zee, and S. Welling-Wester, "Prediction of sequential antigenic regions in proteins," *FEBS Letters*, vol. 188, no. 2, pp. 215–218, 1985.

[17] A. J. P. Alix, "Predictive estimation of protein linear epitopes by using the program PEOPLE," *Vaccine*, vol. 18, no. 3-4, pp. 311–314, 1999.

[18] M. Odorico and J. Pellequer, "BEPITOPE: predicting the location of continuous epitopes and patterns in proteins," *Journal of Molecular Recognition*, vol. 16, no. 1, pp. 20–22, 2003.

[19] S. Saha and G. P. S. Raghava, "BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties," in *ICARIS 2004*, G. Nicosia, V. Cutello, P. J. Bentley, and J. Timis, Eds., vol. 3239 of *Lecture Notes in Computer Science*, pp. 197–204, Springer, Berlin, Germany, 2004.

[20] J. Larsen, O. Lund, and M. Nielsen, "Improved method for predicting linear B-cell epitopes," *Immunome Research*, vol. 2, article 2, 2006.

[21] J. Chen, H. Liu, J. Yang, and K.-C. Chou, "Prediction of linear B-cell epitopes using amino acid pair antigenicity scale," *Amino Acids*, vol. 33, no. 3, pp. 423–428, 2007.

[22] S. Saha and G. P. S. Raghava, "Prediction of continuous B-cell epitopes in an antigen using recurrent neural network," *Proteins*, vol. 65, no. 1, pp. 40–48, 2006.

[23] J. Söllner and B. Mayer, "Machine learning approaches for prediction of linear B-cell epitopes on proteins," *Journal of Molecular Recognition*, vol. 19, no. 3, pp. 200–208, 2006.

[24] Y. El-Manzalawy, D. Dobbs, and V. Honavar, "Predicting linear B-cell epitopes using string kernels," *Journal of Molecular Recognition*, vol. 21, no. 4, pp. 243–255, 2008.

[25] P. Sun, W. Chen, Y. Huang, H. Wang, Z. Ma, and Y. Lv, "Epitope prediction based on random peptide library screening: benchmark dataset and prediction tools evaluation," *Molecules*, vol. 16, no. 6, pp. 4971–4993, 2011.

[26] J. A. Greenbaum, P. H. Andersen, M. Blythe et al., "Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools," *Journal of Molecular Recognition*, vol. 20, no. 2, pp. 75–82, 2007.

[27] Y. Feng, F. Jacobs, E. van Craeyveld et al., "The impact of antigen expression in antigen-presenting cells on humoral immune responses against the transgene product," *Gene Therapy*, vol. 17, no. 2, pp. 288–293, 2009.

[28] M. J. Blythe and D. R. Flower, "Benchmarking B cell epitope prediction: underperformance of existing methods," *Protein Science*, vol. 14, no. 1, pp. 246–248, 2005.

[29] J. V. Ponomarenko and P. E. Bourne, "Antibody-protein interactions: benchmark datasets and prediction tools evaluation," *BMC Structural Biology*, vol. 7, no. 2, article 64, 2007.

[30] A. S. Kolaskar and U. Kulkarni-Kale, "Prediction of three-dimensional structure and mapping of conformational epitopes of envelope glycoprotein of Japanese encephalitis virus," *Virology*, vol. 261, no. 1, pp. 31–42, 1999.

[31] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

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

[33] R. Vita, L. Zarebski, J. A. Greenbaum et al., "The immune epitope database 2.0," *Nucleic Acids Research*, vol. 38, no. 1, pp. D854–D862, 2009.

[34] J. Ponomarenko, N. Papangelopoulos, D. M. Zajonc, B. Peters, A. Sette, and P. E. Bourne, "IEDB-3D: structural data within the immune epitope database," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1164–D1170, 2011.

[35] B. T. Korber, C. Brander, B. F. Haynes et al., *HIV Molecular Immunology Database*, 1998.

[36] U. Kulkarni-Kale, S. Bhosle, and A. S. Kolaskar, "CEP: a conformational epitope prediction server," *Nucleic Acids Research*, vol. 33, no. 2, pp. W168–W171, 2005.

[37] P. H. Andersen, M. Nielsen, and O. Lund, "Prediction of residues in discontinuous B-cell epitopes using protein 3D structures," *Protein Science*, vol. 15, no. 11, pp. 2558–2567, 2006.

[38] R. Rapberger, A. Lukas, and B. Mayer, "Identification of discontinuous antigenic determinants on proteins based on shape complementarities," *Journal of Molecular Recognition*, vol. 20, no. 2, pp. 113–121, 2007.

[39] J. V. Ponomarenko and P. E. Bourne, "Antibody-protein interactions: benchmark datasets and prediction tools evaluation," *BMC Structural Biology*, vol. 7, article 64, 2007.

[40] J. Ponomarenko, H. Bui, W. Li et al., "ElliPro: a new structure-based tool for the prediction of antibody epitopes," *BMC Bioinformatics*, vol. 9, article 514, 2008.

[41] M. J. Sweredoski and P. Baldi, "PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure," *Bioinformatics*, vol. 24, no. 12, pp. 1459–1460, 2008.

[42] V. Moreau, C. Fleury, D. Piquer et al., "PEPOP: computational design of immunogenic peptides," *BMC Bioinformatics*, vol. 9, article 71, 2008.

[43] J. M. Thornton, M. S. Edwards, W. R. Taylor, and D. J. Barlow, "Location of "continuous" antigenic determinants in the protruding regions of proteins," *The EMBO Journal*, vol. 5, no. 2, pp. 409–413, 1986.

[44] N. W. B. Eswar, M. A. Marti-Renom, M. S. Madhusudhan et al., "Comparative protein structure modeling with Modeller," in *Current Protocols in Bioinformatics*, 5. 6. 1–5. 6. 30, Supplement 15, John Wiley & Sons, New York, NY, USA, 2006.

[45] J. Sun, D. Wu, T. Xu et al., "SEPPA: a computational server for spatial epitope prediction of protein antigens," *Nucleic Acids Research*, vol. 37, supplement 2, pp. W612–W616, 2009.

[46] N. D. Rubinstein, I. Mayrose, E. Martz, and T. Pupko, "Epitopia: a web-server for predicting B-cell epitopes," *BMC Bioinformatics*, vol. 10, article 287, 2009.

[47] S. Liang, D. Zheng, C. Zhang, and M. Zacharias, "Prediction of antigenic epitopes on protein surfaces by consensus scoring," *BMC Bioinformatics*, vol. 10, article 302, 2009.

[48] P. Sun, W. Chen, X. Wang, B. Liu, and Y. Lv, "Prediction of antigen epitopes on protein surfaces based on support vector machine," *Advanced Materials Research*, vol. 393-395, pp. 884–889, 2012.

[49] S. Soga, D. Kuroda, H. Shirai, M. Kobori, and N. Hirayama, "Use of amino acid composition to predict epitope residues of individual antibodies," *Protein Engineering, Design and Selection*, vol. 23, no. 6, pp. 441–448, 2010.

[50] J. Sun, T. Xu, S. Wang, G. Li, and D. Wu, "Does difference exist between epitope and non-epitope residues? Analysis of the physicochemical and structural properties on conformational epitopes from B-cell protein antigens," *Immunome Research*, vol. 7, no. 3, pp. 1–11, 2011.

[51] S. Liang, D. Zheng, D. M. Standley, B. Yao, M. Zacharias, and C. Zhang, "EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results," *BMC Bioinformatics*, vol. 11, article 381, 2010.

[52] W. Zhang, Y. Xiong, M. Zhao, H. Zou, X. Ye, and J. Liu, "Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature," *BMC Bioinformatics*, vol. 12, article 341, 2011.

[53] B. Yao, D. Zheng, S. Liang, and C. Zhang, "Conformational B-cell epitope prediction on antigen protein structures: a review of current algorithms and comparison with common binding site prediction methods," *PLoS ONE*, vol. 8, no. 4, Article ID e62249.

[54] O. V. Tsodikov, M. Thomas Record Jr., and Y. V. Sergeev, "Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature," *Journal of Computational Chemistry*, vol. 23, no. 6, pp. 600–609, 2002.

[55] S. J. Hubbard, *NACCESS Computer Program*, University College London, London, UK, 1993.

[56] H. M. Geysen, S. J. Rodda, and T. J. Mason, "A priori delineation of a peptide which mimics a discontinuous antigenic determinant," *Molecular Immunology*, vol. 23, no. 7, pp. 709–715, 1986.

[57] V. Moreau, C. Granier, S. Villard, D. Laune, and F. Molina, "Discontinuous epitope prediction based on mimotope analysis," *Bioinformatics*, vol. 22, no. 9, pp. 1088–1095, 2006.

[58] V. P. Valuev, D. A. Afonnikov, M. P. Ponomarenko, L. Milanesi, and N. A. Kolchanov, "ASPD (Artificially Selected Proteins/Peptides Database): a database of proteins and peptides evolved in vitro," *Nucleic Acids Research*, vol. 30, no. 1, pp. 200–202, 2002.

[59] S. Mandava, L. Makowski, S. Devarapalli, J. Uzubell, and D. J. Rodi, "RELIC—a bioinformatics server for combinatorial peptide analysis and identification of protein-ligand interaction sites," *Proteomics*, vol. 4, no. 5, pp. 1439–1460, 2004.

[60] T. Shtatland, D. Guettler, M. Kossodo, M. Pivovarov, and R. Weissleder, "PepBank—a database of peptides based on sequence text mining and public peptide data sources," *BMC Bioinformatics*, vol. 8, article 280, 2007.

[61] B. Ru, J. Huang, P. Dai et al., "MimoDB: a new repository for mimotope data derived from phage display technology," *Molecules*, vol. 15, no. 11, pp. 8279–8288, 2010.

[62] J. Huang, B. Ru, P. Zhu et al., "MimoDB 2. 0: a mimotope database and beyond," *Nucleic Acids Research*, vol. 40, no. 1, pp. 271–277, 2011.

[63] J. Huang, B. Ru, and P. Dai, "Bioinformatics resources and tools for phage display," *Molecules*, vol. 16, no. 1, pp. 694–709, 2011.

[64] B. M. Mumey, B. W. Bailey, B. Kirkpatrick, A. J. Jesaitis, T. Angel, and E. A. Dratz, "A new method for mapping discontinuous antibody epitopes to reveal structural features of proteins," *Journal of Computational Biology*, vol. 10, no. 3-4, pp. 555–567, 2003.

[65] B. Mumey, N. Ohler, T. Angel et al., "Filtering epitope alignments to improve protein surface prediction," in *Frontiers of High Performance Computing and Networking—ISPA, 2006 Workshops*, G. Min, B. Di Martino, L. Yang, M. Guo, and G. Ruenger, Eds., vol. 4331, pp. 648–657, Springer, Berlin, Germany, 2006.

[66] V. Moreau, C. Granier, S. Villard, D. Laune, and F. Molina, "Discontinuous epitope prediction based on mimotope analysis," *Bioinformatics*, vol. 22, no. 9, pp. 1088–1095, 2006.

[67] T. Castrignanò, P. D. de Meo, D. Carrabino, M. Orsini, M. Floris, and A. Tramontano, "The MEPS server for identifying protein conformational epitopes," *BMC Bioinformatics*, vol. 8, supplement1, article S6, 2007.

[68] A. Schreiber, M. Humbert, A. Benz, and U. Dietrich, "3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins," *Journal of Computational Chemistry*, vol. 26, no. 9, pp. 879–887, 2005.

[69] D. Enshell-Seijffers, D. Denisov, B. Groisman et al., "The mapping and reconstitution of a conformational discontinuous B-cell epitope of HIV-1," *Journal of Molecular Biology*, vol. 334, no. 1, pp. 87–101, 2003.

[70] E. M. Bublil, N. T. Freund, I. Mayrose et al., "Stepwise prediction of conformational discontinuous B-cell epitopes using the mapitope algorithm," *Proteins*, vol. 68, no. 1, pp. 294–304, 2007.

[71] G. F. Denisova, D. A. Denisov, J. Yeung, M. B. Loeb, M. S. Diamond, and J. L. Bramson, "A novel computer algorithm improves antibody epitope prediction using affinity-selected mimotopes: a case study using monoclonal antibodies against the West Nile virus E protein," *Molecular Immunology*, vol. 46, no. 1, pp. 125–134, 2008.

[72] D. A. Denisov, G. F. Denisova, A. Lelic, M. B. Loeb, and J. L. Bramson, "Deciphering epitope specificities within polyserum using affinity selection of random peptides and a novel algorithm based on pattern recognition theory," *Molecular Immunology*, vol. 46, no. 3, pp. 429–436, 2009.

[73] G. F. Denisova, D. A. Denisov, and J. L. Bramson, "Applying bioinformatics for antibody epitope prediction using affinity-selected mimotopes—relevance for vaccine design," *Immunome Research*, vol. 6, supplement 2, article S6, 2010.

[74] I. Halperin, H. Wolfson, and R. Nussinov, "SiteLight: binding-site prediction using phage display libraries," *Protein Science*, vol. 12, no. 7, pp. 1344–1359, 2003.

[75] S. S. Negi and W. Braun, "Automated detection of conformational epitopes using phage display peptide sequences," *Bioinformatics and Biology Insights*, vol. 3, pp. 71–81, 2009.

[76] I. Mayrose, T. Shlomi, N. D. Rubinstein et al., "Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm," *Nucleic Acids Research*, vol. 35, no. 1, pp. 69–78, 2007.

[77] Y. X. Huang, Y. L. Bao, S. Y. Guo, Y. Wang, C. G. Zhou, and Y. X. Li, "Pep-3D-search: a method for B-cell epitope prediction based on mimotope analysis," *BMC Bioinformatics*, vol. 9, article 538, 2008.

[78] W. H. Chen, P. P. Sun, Y. Lu, W. W. Guo, Y. X. Huang, and Z. Q. Ma, "MimoPro: a more efficient Web-based tool for epitope prediction using phage display libraries," *BMC Bioinformatics*, vol. 12, article 199, 2011.

[79] I. Mayrose, O. Penn, E. Erez et al., "Pepitope: epitope mapping from affinity-selected peptides," *Bioinformatics*, vol. 23, no. 23, pp. 3244–3246, 2007.

[80] W. Chen, W. W. Guo, Y. Huang, and Z. Ma, "PepMapper: a collaborative web tool for mapping epitopes from affinity-selected peptides," *PLoS ONE*, vol. 7, no. 5, Article ID e37869, 2012.

[81] H. R. Ansari and G. P. Raghava, "Identification of conformational B-cell epitopes in an antigen from its primary sequence," *Immunome Research*, vol. 6, no. 1, article 6, 2010.

[82] J. Gao, E. Faraggi, Y. Zhou, J. Ruan, and L. Kurgan, "BEST: improved prediction of B-cell epitopes from antigen sequences," *PLoS ONE*, vol. 7, no. 6, Article ID e40104, 2012.

[83] W. Zhang, Y. Niu, Y. Xiong, M. Zhao, R. Yu, and J. Liu, "Computational prediction of conformational B-cell epitopes from antigen primary structures by ensemble learning," *PLoS ONE*, vol. 7, no. 8, Article ID e43575, 2012.

[84] H. Neuvirth, R. Raz, and G. Schreiber, "ProMate: a structure based prediction program to identify the location of protein-protein binding sites," *Journal of Molecular Biology*, vol. 338, no. 1, pp. 181–199, 2004.

[85] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal, "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids," *Nucleic Acids Research*, vol. 38, no. 2, pp. W529–W533, 2010.

[86] S. Liang, C. Zhang, S. Liu, and Y. Zhou, "Protein binding site prediction using an empirical scoring function," *Nucleic Acids Research*, vol. 34, no. 13, pp. 3698–3707, 2006.

[87] I. Kufareva, L. Budagyan, E. Raush, M. Totrov, and R. Abagyan, "PIER: protein interface recognition for structural proteomics," *Proteins*, vol. 67, no. 2, pp. 400–417, 2007.

[88] J. Huang, A. Gutteridge, W. Honda, and M. Kanehisa, "MIMOX: a web tool for phage display based epitope mapping," *BMC Bioinformatics*, vol. 7, article 451, 10 pages, 2006.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 983829, 6 pages
http://dx.doi.org/10.1155/2013/983829

Research Article

Epitope Mapping of Metuximab on CD147 Using Phage Display and Molecular Docking

Bifang He,[1] Canquan Mao,[2] Beibei Ru,[1] Hesong Han,[2] Peng Zhou,[1] and Jian Huang[1]

[1] *Center of Bioinformatics (COBI), Key Laboratory for NeuroInformation of Ministry of Education, University of Electronic Science and Technology of China, Chengdu 610054, China*
[2] *School of Life Science and Engineering, Southwest Jiaotong University, Chengdu 610031, China*

Correspondence should be addressed to Jian Huang; hj@uestc.edu.cn

Received 25 March 2013; Accepted 7 May 2013

Academic Editor: Yanxin Huang

Metuximab is the generic name of Licartin, a new drug for radioimmunotherapy of hepatocellular carcinoma. Although it is known to be a mouse monoclonal antibody against CD147, the complete epitope mediating the binding of metuximab to CD147 remains unknown. We panned the Ph.D.-12 phage display peptide library against metuximab and got six mimotopes. The following bioinformatics analysis based on mimotopes suggested that metuximab recognizes a conformational epitope composed of more than 20 residues. The residues of its epitope may include T28, V30, K36, L38, K57, F74, D77, S78, D79, D80, Q81, G83, S86, N98, Q100, L101, H102, G103, P104, V131, P132, and K191. The homology modeling of metuximab and the docking of CD147 to metuximab were also performed. Based on the top one docking model, the epitope was predicted to contain 28 residues: AGTVFTTV (23–30), I37, D45, E84, V88, EPMGTANIQLH (92–102), VPP (131–133), Q164, and K191. Almost half of the residues predicted on the basis of mimotope analysis also appear in the docking result, indicating that both results are reliable. As the predicted epitopes of metuximab largely overlap with interfaces of CD147-CD147 interactions, a structural mechanism of metuximab is proposed as blocking the formation of CD147 dimer.

1. Introduction

Metuximab is the generic name of HAb18, a mouse monoclonal antibody of IgG1 class developed by Chen et al. in 1989 [1]. The hybridoma producing HAb18 was made from mice immunized with a cell suspension of fresh human hepatocellular carcinoma tissues. The antigen recognized by HAb18 was accordingly called HAb18G, which was later identified as an isoform of basigin [2]. The products of basigin gene have many well-known names, for example, cluster of differentiation 147 (CD147), extracellular matrix metalloproteinase inducer (EMMPRIN), and so on. In this paper, we use CD147 hereafter to refer to the antigen recognized by metuximab.

CD147 is a transmembrane glycoprotein of the immunoglobulin superfamily. It has been involved in various physiological functions such as spermatogenesis [3], embryo implantation [4], tissue remodeling [5], and diverse pathological processes such as neuroinflammation [6], Alzheimer's disease [7], malaria infection [8], and tumor progression [9]. Though widely expressed on numerous cell types, CD147 is highly enriched on the surface of cancer cells, especially on those of epithelial origin, for example, breast cancer [10] and liver cancer [11]. Thus, it has been taken as a biomarker that can be used in cancer detection [12]. Furthermore, CD147 has also been proposed to be a new drug target for developing therapeutics against inflammation, malaria [13], and cancer [14]. Metuximab is a success case which targets CD147. The iodine-131-labeled F(ab′)2 fragment of metuximab has been reported to be safe and effective for targeted treatment of hepatocellular carcinoma in clinical trials [15, 16]. The injection with the brand name Licartin was approved as a new drug for radioimmunotherapy of hepatocellular carcinoma by the State Food and Drug Administration, China, in 2005.

Where does metuximab bind to CD147? The answer will help us understand the mechanism of CD147 function and benefit the development of new drugs targeting CD147. Using binding assays to a series of truncated fragments of CD147 ectodomain, Ku et al. reported that the segment of

39LTCSLNDSATEV50 was the epitope on CD147 recognized by metuximab [17]. Yu et al. docked the Fv fragment of metuximab onto the N-terminal domain of CD147 in a head-to-head manner when the crystal structure of CD147 ectodomain was resolved [18]. Their model and experiment results suggested the residues E49, T51, and D65 on CD147 might also play an important role in the interaction between CD147 and metuximab [18]. However, a panorama of the epitope mediating the binding of metuximab to CD147 has not been proposed yet.

Crystallographic analysis of antigen-antibody complex is the most accurate approach to mapping an epitope. However, it is time consuming and sometimes technically difficult or even impossible to get an antigen-antibody complex crystallized. As an alternative choice for epitope mapping with lower but acceptable precision, mimotope analysis is becoming to be an increasingly popular method for its cheapness and quickness [19, 20]. In this study, the epitope of metuximab was defined completely at the residue level using phage display and the following bioinformatics analysis. The result was then validated using molecular modeling and docking. The panoramic model where CD147 is recognized by metuximab will provide valuable information and better structural basis for decoding CD147 and developing relevant drugs.

2. Materials and Methods

2.1. Biopanning of a 12-Mer Phage Display Library. The F(ab′)2 fragment of metuximab as freeze-dried powder with purity above 97% was provided by Chengdu Huasun Bio-Tech Co. LTD. The Ph.D.-12 phage display peptide library that displays 2.7×10^9 unique 12-amino acid peptides fused to the pIII minor coat protein of the M13 filamentous phage was purchased from New England BioLabs. Three successive rounds of biopannings were performed with the F(ab′)2 fragment of metuximab as the capture reagent coated on 96-well microtiter plates, as described in the manufacturer's manual with modifications.

In brief, the sample was diluted to a concentration of 200 μg/mL in 0.1 M NaHCO3 (pH 8.6). 100 μL of the above solution was transferred to a 96-well plate and incubated overnight at 4°C with gentle agitation in a humidified container. Wells coated with buffer were used as the negative control. The coating solution was removed from the wells, and the plate was tapped onto a clean paper towel to remove residual solution. Each well was filled with the blocking buffer (0.1 M sodium bicarbonate buffer pH 8.6 plus 5 mg/mL BSA). The plate was incubated for 1 hour at 4°C. The blocking solution was removed, and the plate was tapped onto a clean paper towel to remove residual solution. Each well was washed rapidly 6 times with 1% TBST (Tris-buffered saline with 0.1% Tween 20). The plate was swirled repeatedly when coating and washing to ensure that each well including its sides was coated and washed completely. 2×10^{11} phages from the library were mixed with 100 μL TBST and transferred onto coated well. The plate was incubated at room temperature for 45 min with gentle rocking. Then, the supernatant was removed and the plate was washed ten times with 200 μL TBST buffer. The bound phages were eluted with 100 μL 0.2 M glycine-HCl (pH 2.2) plus 1 mg/mL of BSA. The eluate was transferred to an Eppendorf tube and neutralized immediately by adding 30 μL 1 M Tris-HCl (PH9.1). After titration, the eluate was amplified in *E. coli* strain ER2738 culture for additional two pannings using the same PFUs of total phage in each round. After the third panning experiment, the final eluate was mixed with ER2738 host cells, diluted, and spread on LB-Xgal/IPTG plates. Twenty isolated plaques were randomly picked and amplified for DNA sequencing.

2.2. Epitope Mapping Based on Mimotope Analysis. The peptides displayed on the selected phages were deduced from the results of DNA sequencing. The data was firstly cleaned using the tools in the SAROTUP suite to exclude any possible target-unrelated peptides [21–23]. The left peptides were then mapped back to the surface of CD147 based on its crystal structure (PDB: 3B5H) using the EpiSearch program by default parameters [24]. The mapping result of each peptide was united to make the epitope on CD147 recognized by metuximab.

2.3. Molecular Modeling and Docking. The sequences of the variable heavy chain (VH) and light chain (VL) of metuximab were extracted from the United State patent with the number US7638619 [25]. The corresponding GenBank accession numbers of VH and VL are ADC21949.1 and ADC21950.1, respectively. The sequences were manually checked. Only segments 20–136 of VH and 21–130 of VL were submitted to the RosettaAntibody server to construct the model for variable domain (Fv) of metuximab [26]. For each framework region and complementarity determining region (CDR) of metuximab, the best templates were used for homology modeling (see Table 1).

As shown in Table 1, the H3 loop of metuximab does not have any sequence match. It was thus modeled ab initio by the RosettaAntibody server [26]. Top ten models with energy minimized were given by the server, and the top one (see Figure 1) was used as the receptor for docking.

The 3D structure of CD147 monomer (3B5H, chain C) was then docked to the Fv model of metuximab using the program ZDOCK with framework region blocked [27]. The results were evaluated using ZRANK and optimized using RDOCK [28, 29]. The top one docking model from RDOCK was picked up as the theoretical metuximab-CD147 complex. The interfaces between metuximab and CD147 were then computed with the program PISA [30]. The one on CD147 side was taken as the epitope produced by molecular docking.

3. Results and Discussion

3.1. Analysis on Panning Results. The titer of the eluate after each round of pannings increased from 10^3, 10^4 to 10^6 PFU/mL, indicating an efficient enrichment of phages specifically binding to metuximab. After the third panning, twenty phage clones were randomly picked and sent to DNA sequencing. The deduced amino acids sequences are listed in Table 2.

Table 1: Templates used in modeling metuximab.

Region*	PDB	Identity	Align length	Mismatches	Gap openings	E value
HFR	1SBS	95.52	67	3	0	$0.00E+00$
LFR	1IAI	85.48	62	9	0	$3.00E-22$
H1	2DLF	100	10	0	0	$1.00E-07$
H2	2DLF	75	20	3	2	$3.00E-08$
H3	1A0Q	0	0	0	0	$0.00E+00$
L1	1QNZ	100	6	0	0	$5.80E-02$
L2	3FCT	83.33	6	1	0	$3.30E-01$
L3	1NCD	77.78	9	2	0	$5.00E-03$

*HFR: heavy chain framework region; LFR: light chain framework region; H1, H2, H3, L1, L2, and L3 refer to CDR1, CDR2, and CDR3 loops of heavy and light chain, respectively.

Table 2: Peptides selected from Ph.D.-12 phage display peptide library using metuximab.

Number	Sequence	Occurrence
P1	YPHFHKHTLRGH	9
P2	YPHFHKHSLRGQ	1
P3	DHKPFKPTHRTL	1
P4	FHKPFKPTHRTL	1
P5	QSSCHKHSVRGR	1
P6	QSSFSNHSVRRR	1
P7	DFDVSFLSARMR	6

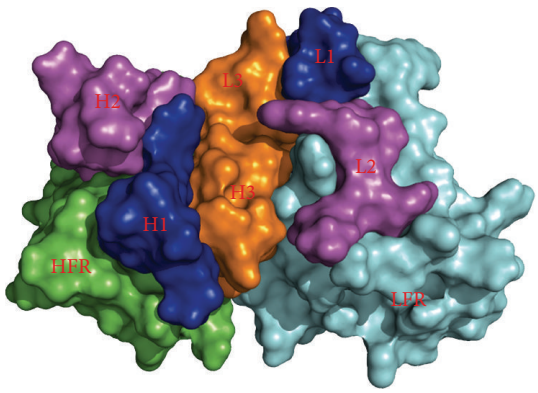


Figure 1: Fv model of metuximab.

As shown in Table 2, seven unique sequences were obtained from biopannings. Among them, the peptide YPHFHKHTLRGH is most frequent. Just by visual inspection, these peptides can be grouped into 4 clusters, that is, P1 and P2, P3 and P4, P5 and P6, and P7.

3.2. Epitope Mapping Results from Mimotope Analysis. These peptides were checked using tools in the SAROTUP suite. Interestingly, it was reported that the peptide DFDVSFLSARMR had also been panned out from the Ph.D.-12 phage display peptide library using the protein tonB of *E. coli* [31]. To avoid any possible TUP, this peptide was dropped from the following epitope mapping based on mimotopes. The left peptides were then used together as inputs of the EpiSearch program [24]. By default parameters, they were mapped back to the surface of CD147 monomer based on the crystal structure (PDB: 3B5H, Chain C). The analysis found two solutions centered at residue Lys191 (score = 0.780) and Leu101 (score = 0.713), respectively. For each solution, the high-scoring patch for each peptide was united to make the whole epitope predicted to be recognized by metuximab.

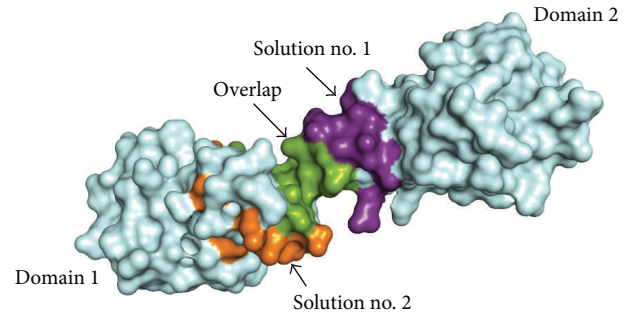


Figure 2: Epitope mapping results from mimotope analysis. Residues only appearing in solutions nos. 1 and 2 are colored in purple and orange, respectively; the overlapping region of the two solutions is drawn in green. All other parts of CD147 surface are presented in pale cyan.

As shown in Tables 3 and 4, the two epitopes predicted by EpiSearch are similar in size and residue composition as well. For example, both of them contain 22 residues in total. Furthermore, 11 residues, printed in italics in Tables 3 and 4, are identical in two solutions, indicating a consistence in some degree between the two solutions. As shown in Figure 2, most residues of solution no. 1 locate on the domain 2 of CD147 and the main part of solution no. 2 is on the surface of the domain 1. The two solutions overlap at the area between the two domains of CD147.

When the parameter "Accuracy cutoff" (allowed mismatch) was set from 3 (default) to 2 (a stricter value), only solution 2 was left. Thus, solution 2 might be a more accurate prediction, and we considered this solution to be the epitope predicted by phage display and mimotope analysis.

It has been reported that the frequencies of peptides are not correlated to their binding strength and the diversity of peptides are important in analysis [32]. Therefore, all peptides

Table 3: Epitope mapping solution no. 1 based on mimotope analysis.

Mimotope	Predicted epitopic residues
YPHFHKHTLRGH	L101, H102, G103, P104, R106, P132, P133, T135, T188, K191, G192
YPHFHKHSLRGQ	S78, Q81, Q100, L101, H102, G103, P104, R106, S128, S130, P132, P133, S189, S190, K191, G192, S193
DHKPFKPTHRTL	L101, H102, P104, R106, P132, P133, T135, T188, K191, D194
FHKPFKPTHRTL	L101, H102, P104, R106, P132, P133, T135, T188, K191
QSSCHKHSVRGR	V30, S78, Q81, Q100, H102, G103, R106, S128, S130, V131, S189, S190, K191, G192, S193
QSSFSNHSVRRR	V30, S78, Q81, Q100, H102, R106, S128, S130, V131, S189, S190, S193
Union	*V30*, *S78*, *Q81*, *Q100*, *L101*, *H102*, *G103*, *P104*, **R106**, **S128**, **S130**, *V131*, *P132*, **P133**, **T135**, **T188**, **S189**, **S190**, *K191*, **G192**, **S193**, **D194**; 22 residues in total.

Table 4: Epitope mapping solution no. 2 based on mimotope analysis.

Mimotope	Predicted epitopic residues
YPHFHKHTLRGH	T28, K36, L38, K57, F74, G83, L101, H102, G103, P104, P132, K191
YPHFHKHSLRGQ	K36, L38, K57, F74, S78, Q81, G83, S86, Q100, L101, H102, G103, P104, P132, K191
DHKPFKPTHRTL	T28, K36, L38, K57, F74, D77, D79, D80, L101, H102, P104, P132, K191
FHKPFKPTHRTL	T28, K36, L38, K57, F74, L101, H102, P104, P132, K191
QSSCHKHSVRGR	V30, K36, K57, S78, Q81, G83, S86, Q100, H102, G103, V131, K191
QSSFSNHSVRRR	V30, F74, S78, Q81, S86, N98, Q100, H102, V131
Union	**T28**, *V30*, **K36**, **L38**, **K57**, **F74**, **D77**, *S78*, **D79**, **D80**, *Q81*, **G83**, **S86**, **N98**, *Q100*, *L101*, *H102*, *G103*, *P104*, *V131*, *P132*, *K191*; 22 residues in total.

were treated equally in our study, although their occurrences in the panning results were quite different. Indeed, the P1 mimotope YPHFHKHTLRGH, the most frequently appeared peptide, did not contain more epitope residues than others when all these peptides were mapped back to the surface of CD147 and compared with the docking results.

We have also used other tools such as PepMapper [33–35] to interpret the phage display data and got some results similar to EpiSearch. This makes the prediction above even convincing.

3.3. Epitope Mapping Results from Molecular Docking. The computation of the top one model from RDOCK results revealed that metuximab might bind to a conformational epitope. As shown in Figure 3, the epitope was predicted to contain 28 residues: 23AGTVF**T**T**V**30, I37, D45, E84, V88, 92EPMGTA**NI****QLH**102, 131**VP**P133, Q164, and **K**191.

Other nine models of the top ten RDOCK poses were also used to compute the theoretical epitopes. The results show that seven of them were identical to that of top one model. The epitope derived from the 7th model is a little bit different. Besides including all residues shown in Figure 3, it also contains residues N44, T48, and F89. Only the epitope derived from the 5th model is quite different from that of the top one model. Since the computational results of top ten models are quite consistent with top one, we believe the epitope shown in Figure 3 is reasonable.

Though there are arguments about the role of molecular docking in epitope mapping, the results of molecular docking can be very accurate. In 2001, Saphire et al. solved the structure of b12, a neutralizing human IgG against HIV-1. They docked gp120 to b12 and predicted the epitope recognized by b12 [36]. Six years later, the crystal structure of gp120 and b12 complex was solved and the true epitope was found to be nearly the same as the prediction [37]. As the structure of metuximab-CD147 complex is not solved yet, the accuracy of our docking results cannot be validated at present. However, when the results of mimotope analysis and molecular docking were compared, a significant overlap was found. For example, almost half residues of the phage display solution no. 2 also appear in the docking result (see the residues in bold in Section 3.3). The consistency between phage display and molecular docking suggests that both predictions are reliable.

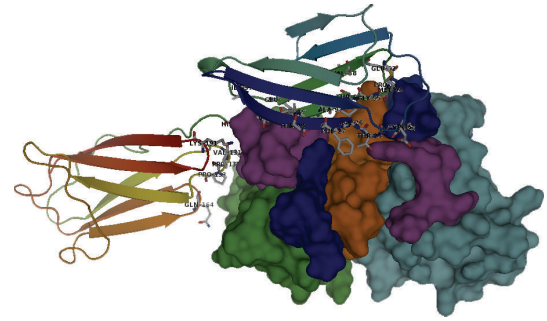


Figure 3: Theoretical model of metuximab-CD147 complex. CD147 was displayed as cartoon in rainbow color, and the interface residues on it were drawn as sticks and labeled.

3.4. Structural Insights into Metuximab Mechanism. CD147 exists in several forms, such as monomer, dimer, and polymer [18]. It has been observed that CD147 can bind to soluble

CD147 [38]. Very recently, Cui et al. found that dimerization was essential for CD147 to promote tumor invasion via MAPK pathway [39]. According to the solved crystal structure of CD147 (3B5H), the interactions between the four monomer chains (i.e., chain A, B, C, and D) of CD147 might represent four possible ways to form a CD147 dimer [18]. Two CD147 monomers on the membranes of two cells may interact with each other via their N-terminal domain (domain 1) just like forming the AC or BC dimer. A soluble CD147 may also bind to another CD147 on cell membrane through its C-terminal domain (domain 2), which is similar to AD dimer. The interaction between chain D and D′ is deemed a result from crystal packing, which is mainly mediated by 59GVVKEDA66 [18].

Very interestingly, the predicted epitopes of metuximab based on either phage display or molecular docking overlap the interfaces between AC, BC, and AD dimmers. Coincidentally, the nonspecific D D′ dimer is the only exception, which has no intersection with the epitope of metuximab. Thus, at least one mechanism of metuximab is to block CD147-CD147 interactions. Therefore, peptides obtained from screening the phage-displayed random peptide library might also have potential applications in blocking CD147 pathways.

4. Conclusions

According to the results from and analyses on molecular docking and phage display experiments, we conclude that metuximab recognizes a conformational epitope composed of more than 20 residues. These residues mainly locate on the N-terminal domain surface of CD147 and largely overlap with interfaces of CD147-CD147 interactions. Blocking the formation of CD147-CD147 dimer may be an important mechanism of metuximab function.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper. This work was supported in part by the National Natural Science Foundation of China under Grant 61071177 and the Program for New Century Excellent Talents in University (NCET-12-0088).

References

[1] Z. N. Chen, Y. F. Liu, and J. Z. Yang, "Production of a monoclonal antibody against human hepatocellular carcinoma and locating the corresponding antigen P60 using immunohistochemistry," *Dan Ke Long Kang Ti Tong Xun*, no. 2, pp. 33–36, 1989.

[2] Z. N. Chen, Z. Yang, L. Mi, J. L. Jiang, and X. Guo, "Analysis on the structure and function of hepatoma transfer-associated factor HAb18G," *Journal of Cellular and Molecular Immunology*, vol. 15, no. 1, p. 34, 1999.

[3] H. Chen, K. Lam Fok, X. Jiang, and H. C. Chan, "New insights into germ cell migration and survival/apoptosis in spermatogenesis: lessons from CD147," *Spermatogenesis*, vol. 2, no. 4, pp. 264–272, 2012.

[4] L. Chen, R. J. Belton Jr., and R. A. Nowak, "Basigin-mediated gene expression changes in mouse uterine stromal cells during implantation," *Endocrinology*, vol. 150, no. 2, pp. 966–976, 2009.

[5] E. Huet, E. E. Gabison, S. Mourah, and S. Menashi, "Role of emmprin/CD147 in tissue remodeling," *Connective Tissue Research*, vol. 49, no. 3-4, pp. 175–179, 2008.

[6] S. M. Agrawal and V. W. Yong, "The many faces of EMMPRIN—roles in neuroinflammation," *Biochimica et Biophysica Acta*, vol. 1812, no. 2, pp. 213–219, 2011.

[7] L. J. Kanyenda, G. Verdile, S. Boulos et al., "The dynamics of CD147 in Alzheimer's disease development and pathology," *Journal of Alzheimer's Disease*, vol. 26, no. 4, pp. 593–605, 2011.

[8] C. Crosnier, L. Y. Bustamante, S. J. Bartholdson et al., "Basigin is a receptor essential for erythrocyte invasion by Plasmodium falciparum," *Nature*, vol. 480, no. 7378, pp. 534–537, 2011.

[9] T. Kanekura and X. Chen, "CD147/basigin promotes progression of malignant melanoma and other cancers," *Journal of Dermatological Science*, vol. 57, no. 3, pp. 149–154, 2010.

[10] F. Liu, L. Cui, Y. Zhang et al., "Expression of HAb18G is associated with tumor progression and prognosis of breast carcinoma," *Breast Cancer Research and Treatment*, vol. 124, no. 3, pp. 677–688, 2010.

[11] J. Xu, H. Y. Xu, Q. Zhang et al., "HAb18G/CD147 functions in invasion and metastasis of hepatocellular carcinoma," *Molecular Cancer Research*, vol. 5, no. 6, pp. 605–614, 2007.

[12] Y. Li, J. Xu, L. Chen et al., "HAb18G (CD147), a cancer-associated biomarker and its role in cancer detection," *Histopathology*, vol. 54, no. 6, pp. 677–687, 2009.

[13] T. Muramatsu, "Basigin: a multifunctional membrane protein with an emerging role in infections by malaria parasites," *Expert Opinion on Therapeutic Targets*, vol. 16, no. 10, pp. 999–1011, 2012.

[14] U. H. Weidle, W. Scheuer, D. Eggle, S. Klostermann, and H. Stockinger, "Cancer-related issues of CD147," *Cancer Genomics and Proteomics*, vol. 7, no. 3, pp. 157–169, 2010.

[15] J. Xu, Z. Y. Shen, X. G. Chen et al., "A randomized controlled trial of licartin for preventing hepatoma recurrence after liver transplantation," *Hepatology*, vol. 45, no. 2, pp. 269–276, 2007.

[16] Z. N. Chen, L. Mi, J. Xu et al., "Targeting radioimmunotherapy of hepatocellular carcinoma with iodine (131I) metuximab injection: clinical phase I/II trials," *International Journal of Radiation Oncology, Biology, Physics*, vol. 65, no. 2, pp. 435–444, 2006.

[17] X. M. Ku, C. G. Liao, Y. Li et al., "Epitope mapping of series of monoclonal antibodies against the hepatocellular carcinoma-associated antigen HAb18G/CD147," *Scandinavian Journal of Immunology*, vol. 65, no. 5, pp. 435–443, 2007.

[18] X. L. Yu, T. Hu, J. M. Du et al., "Crystal structure of HAb18G/CD147: implications for immunoglobulin superfamily homophilic adhesion," *Journal of Biological Chemistry*, vol. 283, no. 26, pp. 18056–18065, 2008.

[19] J. Huang, B. Ru, and P. Dai, "Prediction of protein interaction sites using mimotope analysis," in *Protein-Protein Interactions—Computational and Experimental Tools*, W. Cai and H. Hong, Eds., pp. 189–206, InTech, 2012.

[20] J. Huang, B. Ru, and P. Dai, "Bioinformatics resources and tools for phage display," *Molecules*, vol. 16, no. 1, pp. 694–709, 2011.

[21] J. Huang, B. Ru, P. Zhu et al., "MimoDB 2.0: a mimotope database and beyond," *Nucleic Acids Research*, vol. 40, Database issue, pp. D271–D277, 2012.

[22] B. Ru, J. Huang, P. Dai et al., "MimoDB: a new repository for mimotope data derived from phage display technology," *Molecules*, vol. 15, no. 11, pp. 8279–8288, 2010.

[23] J. Huang, B. Ru, S. Li, H. Lin, and F. B. Guo, "SAROTUP: scanner and reporter of target-unrelated peptides," *Journal of Biomedicine and Biotechnology*, vol. 2010, Article ID 101932, 7 pages, 2010.

[24] S. S. Negi and W. Braun, "Automated detection of conformational epitopes using phage display peptide sequences," *Bioinformatics and Biology Insights*, vol. 3, pp. 71–81, 2009.

[25] Z. Chen, J. Xing, and S. Zhang, "Variable region gene of heavy/light chain of anti-human hepatoma monoclonal antibody HAb 18 and use thereof," US7638619B2, 2009.

[26] A. Sircar, E. T. Kim, and J. J. Gray, "RosettaAntibody: antibody variable region homology modeling server," *Nucleic Acids Research*, vol. 37, Web Server issue, pp. W474–W479, 2009.

[27] A. Tovchigrechko and I. A. Vakser, "GRAMM-X public web server for protein-protein docking," *Nucleic Acids Research*, vol. 34, Web Server issue, pp. W310–W314, 2006.

[28] K. Wiehe, B. Pierce, W. W. Tong, H. Hwang, J. Mintseris, and Z. Weng, "The performance of ZDOCK and ZRANK in rounds 6-11 of CAPRI," *Proteins*, vol. 69, no. 4, pp. 719–725, 2007.

[29] K. Wiehe, B. Pierce, J. Mintseris et al., "ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5," *Proteins*, vol. 60, no. 2, pp. 207–213, 2005.

[30] E. Krissinel and K. Henrick, "Inference of macromolecular assemblies from crystalline state," *Journal of Molecular Biology*, vol. 372, no. 3, pp. 774–797, 2007.

[31] D. M. Carter, I. R. Miousse, J. N. Gagnon et al., "Interactions between TonB from Escherichia coli and the periplasmic protein FhuD," *Journal of Biological Chemistry*, vol. 281, no. 46, pp. 35413–35424, 2006.

[32] R. Derda, S. K. Y. Tang, S. C. Li, S. Ng, W. Matochko, and M. R. Jafari, "Diversity of phage-displayed libraries of peptides during panning and amplification," *Molecules*, vol. 16, no. 2, pp. 1776–1803, 2011.

[33] W. Chen, W. W. Guo, Y. Huang, and Z. Ma, "PepMapper: a collaborative web tool for mapping epitopes from affinity-selected peptides," *PLoS One*, vol. 7, no. 5, Article ID e37869, 2012.

[34] W. H. Chen, P. P. Sun, Y. Lu, W. W. Guo, Y. X. Huang, and Z. Q. Ma, "MimoPro: a more efficient Web-based tool for epitope prediction using phage display libraries," *BMC Bioinformatics*, vol. 12, article 199, 2011.

[35] Y. X. Huang, Y. L. Bao, S. Y. Guo, Y. Wang, C. G. Zhou, and Y. X. Li, "Pep-3D-Search: a method for B-cell epitope prediction based on mimotope analysis," *BMC Bioinformatics*, vol. 9, article 538, 2008.

[36] E. O. Saphire, P. W. H. I. Parren, R. Pantophlet et al., "Crystal structure of a neutralizing human IgG against HIV-1: a template for vaccine design," *Science*, vol. 293, no. 5532, pp. 1155–1159, 2001.

[37] T. Zhou, L. Xu, B. Dey et al., "Structural definition of a conserved neutralization epitope on HIV-1 gp120," *Nature*, vol. 445, no. 7129, pp. 732–737, 2007.

[38] R. J. Belton Jr., L. Chen, F. S. Mesquita, and R. A. Nowak, "Basigin-2 is a cell surface receptor for soluble basigin ligand," *Journal of Biological Chemistry*, vol. 283, no. 26, pp. 17805–17814, 2008.

[39] H. Y. Cui, T. Guo, S. J. Wang et al., "Dimerization is essential for HAb18G/CD147 promoting tumor invasion via MAPK pathway," *Biochemical and Biophysical Research Communications*, vol. 419, no. 3, pp. 517–522, 2012.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 470390, 11 pages
http://dx.doi.org/10.1155/2013/470390

Research Article

Uses of Phage Display in Agriculture: Sequence Analysis and Comparative Modeling of Late Embryogenesis Abundant Client Proteins Suggest Protein-Nucleic Acid Binding Functionality

Rekha Kushwaha,[1,2] A. Bruce Downie,[2,3] and Christina M. Payne[4,5]

[1] *Agricultural Science Center, Department of Horticulture, University of Kentucky, Lexington, KY 40546, USA*
[2] *Seed Biology Group, University of Kentucky, Lexington, KY 40546, USA*
[3] *Plant Science Building, Department of Horticulture, University of Kentucky, Lexington, KY 40546, USA*
[4] *Department of Chemical and Materials Engineering, University of Kentucky, Lexington, KY 40506, USA*
[5] *Center for Computational Sciences, University of Kentucky, Lexington, KY 40506, USA*

Correspondence should be addressed to Christina M. Payne; christy.payne@uky.edu

Received 27 February 2013; Accepted 2 April 2013

Academic Editor: Jian Huang

A group of intrinsically disordered, hydrophilic proteins—Late Embryogenesis Abundant (LEA) proteins—has been linked to survival in plants and animals in periods of stress, putatively through safeguarding enzymatic function and prevention of aggregation in times of dehydration/heat. Yet despite decades of effort, the molecular-level mechanisms defining this protective function remain unknown. A recent effort to understand LEA functionality began with the unique application of phage display, wherein phage display and biopanning over recombinant Seed Maturation Protein homologs from *Arabidopsis thaliana* and *Glycine max* were used to retrieve client proteins at two different temperatures, with one intended to represent heat stress. From this previous study, we identified 21 client proteins for which clones were recovered, sometimes repeatedly. Here, we use sequence analysis and homology modeling of the client proteins to ascertain common sequence and structural properties that may contribute to binding affinity with the protective LEA protein. Our methods uncover what appears to be a predilection for protein-nucleic acid interactions among LEA client proteins, which is suggestive of subcellular residence. The results from this initial computational study will guide future efforts to uncover the protein protective mechanisms during heat stress, potentially leading to phage-display-directed evolution of synthetic LEA molecules.

1. Introduction

Water is essential for life. Despite this apparent truism, there are organisms that have phases of their life cycle during which they can withstand dehydration to less than 5% water content on a fresh weight basis. This phenomenon has become known as "anhydrobiosis" or life without water [1, 2]. One of the means by which those organisms capable of anhydrobiosis are thought to retain viability at very low moisture content is through the vitrification of the cytoplasm upon water removal [3, 4]. The cytoplasmic phase transitions, from liquid to viscous to glass, are thought to increasingly impede deleterious biochemical reactions while progressively dampening respiration [5]. A second requirement is to protect those cellular components, dependent on water to maintain their structure/function, using so-called "water replacement" by specific, non-reducing oligosaccharides [2] which, in conjunction with highly hydrophilic proteins, can also enhance the quality and persistence of the glassy state [6, 7]. A third means is to prevent the aggregation of cellular constituents as water is withdrawn, and the distance between macromolecules diminishes [8, 9]. All of these properties have been assigned to various families of the Late Embryogenesis Abundant (LEA) proteins which were first identified [10] and then named [11] from studies of cotton seed proteins found in the embryo.

The characteristic intrinsically disordered structure and high hydrophilicity of the LEA proteins have been used to argue that they may act in a variety of ways to replace water (or compensate for its loss) in dehydrating tissues [12, 13]. Although there are two known LEA structures [14, 15], many of the proteins belonging to this family are dynamically disordered by design [16–18]. This has reasonably led to difficulties in obtaining structural information despite the use of a variety of techniques [19, 20], temperatures, and additives [17, 21]. Although obtaining crystal structures for most LEAs is not likely in the near future, structures of the preferential LEA client proteins may be estimated through homology modeling [22–24] as the same data allowing client protein identification also permits the identification of the region of the client protein to which the LEAs bind. Understanding which proteins are a particular LEA's preeminent substrates provides insights into those functional processes most at risk for dehydration/thermal damage, suggesting novel ways forward in producing more drought-/heat-resistant species. Identification of hallmarks within the bound regions of LEA client proteins will provide the first clues as to which protein topologies are particularly prone to dehydration/heat damage. We hypothesize the regions require protection, which may be achieved through LEA protein binding.

Here, we report functional insights relative to LEA client proteins from application of sequence analysis and comparative modeling. Our examination focuses on identifying commonalities within the set of 21 putative LEA protein interactions previously identified using phage display [25]. Sequence analysis suggests a common theme among many of the LEA client proteins may be protein-nucleic interaction motifs which may provide clues regarding subcellular residence of the LEA proteins themselves. Homology modeling, where feasible, uncovers several structures, varying both in length and tertiary structure, whose common thread may be related to dynamic and chemical behavior more than structural or sequence similarity.

2. Comparative Modeling Methods

Previously, phage display with *Arabidopsis* seed cDNA libraries in T7 phage was used in biopans of recombinant *Arabidopsis thaliana* seed maturation protein 1 (SMP1) and its *Glycine max* homologue, *Gm*PM28 [25] (LEA proteins). Biopanning was performed at 25°C and 41°C to identify proteins potentially involved in induction of secondary dormancy of *Arabidopsis thaliana* as a result of heat stress (see our companion manuscript for a brief synopsis of seed maturation). Figure 1 illustrates the 21 putative LEA client proteins identified through phage display. The proteins are labeled by the *Arabidopsis* Information Resource (TAIR) locus identifier. Within each plot, the LEA to which the protein binds and the temperature of the biopan are given. These proteins serve as the basis for our sequence analysis and comparative modeling investigation.

The full-length protein sequences to which LEA proteins of the Seed Maturation Protein family bound in phage display [25] were acquired from TAIR. Each protein was used in screens to identify homologs for which suitable three-dimensional (3D) structures had been solved. For the comparative modeling effort, we focused specifically on the regions identified as binding to the LEA homologues. Based on availability of 3D structures similar to these regions, the number of hits was narrowed down to 7 from the original 21 proteins being assessed (AT1G54870.1, AT1G75830.1, AT3G55170.1, AT3G58680.1, AT5G18380.1, AT5G44120.1, and AT5G46430.2).

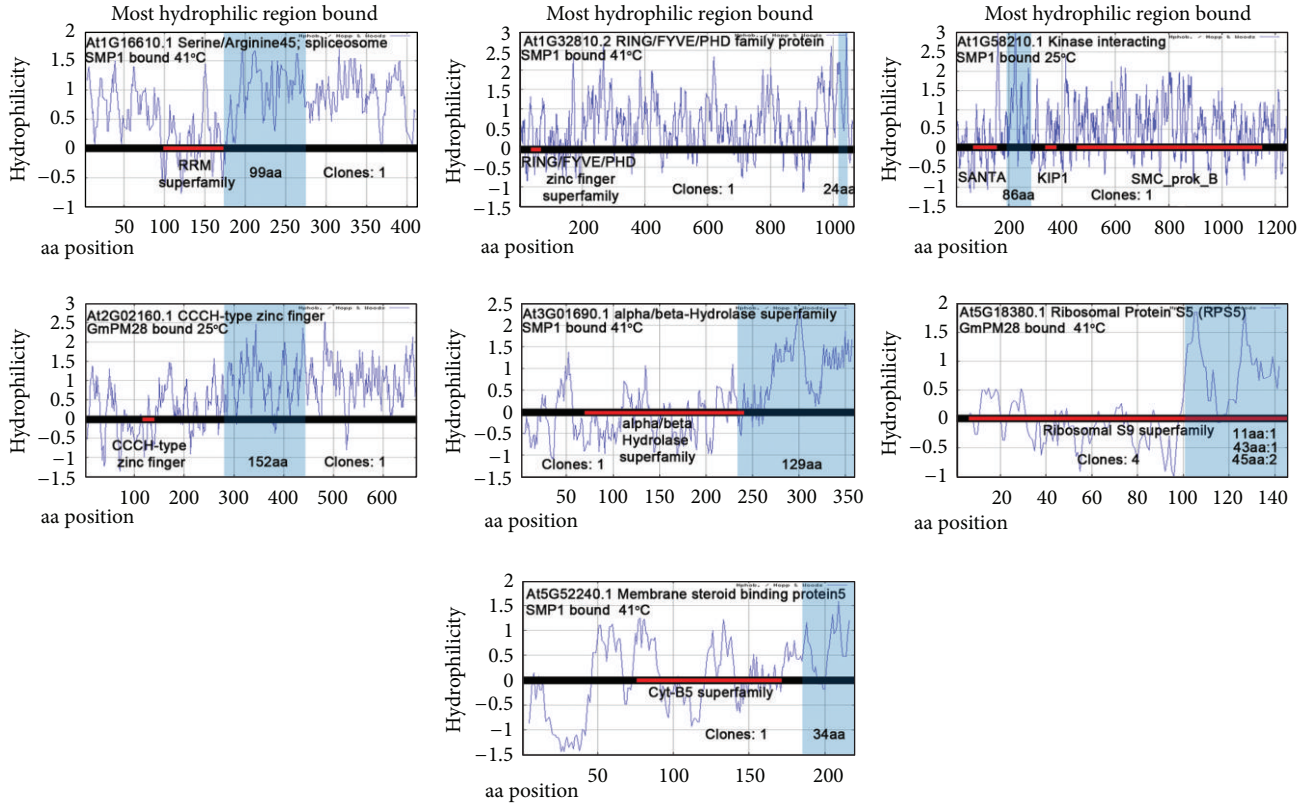
Homology modeling of the seven LEA client proteins was performed using the Bioinformatics Toolkit from the Max-Planck Institute for Developmental Biology [26]. This suite integrates a number of utilities necessary to complete the modeling process. For each of the seven proteins, HHpred was used to predict secondary structure and sequence homology [27, 28]. HHblits was used to build multiple sequence alignments as input to the homology modeling software [29]. Homology modeling was performed using MODELLER [30].

Each protein used different templates for which the atomic coordinates were obtained from the RCSB Protein Data Bank [31]. Table 1 summarizes the templates used along with a brief description of each. For each of the models, the standard automated MODELLER procedure for structure modeling and optimization was used. This includes the initial rule-based determination of spatial restraints from the alignment and optimization through minimization of restraint violations. Several of the homology models generated include segments other than the bound regions of interest; however, for the purposes of this project, we limit discussion to the phage-display-recovered regions of the LEA client proteins. Visualization of the proteins and templates was accomplished with PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.).

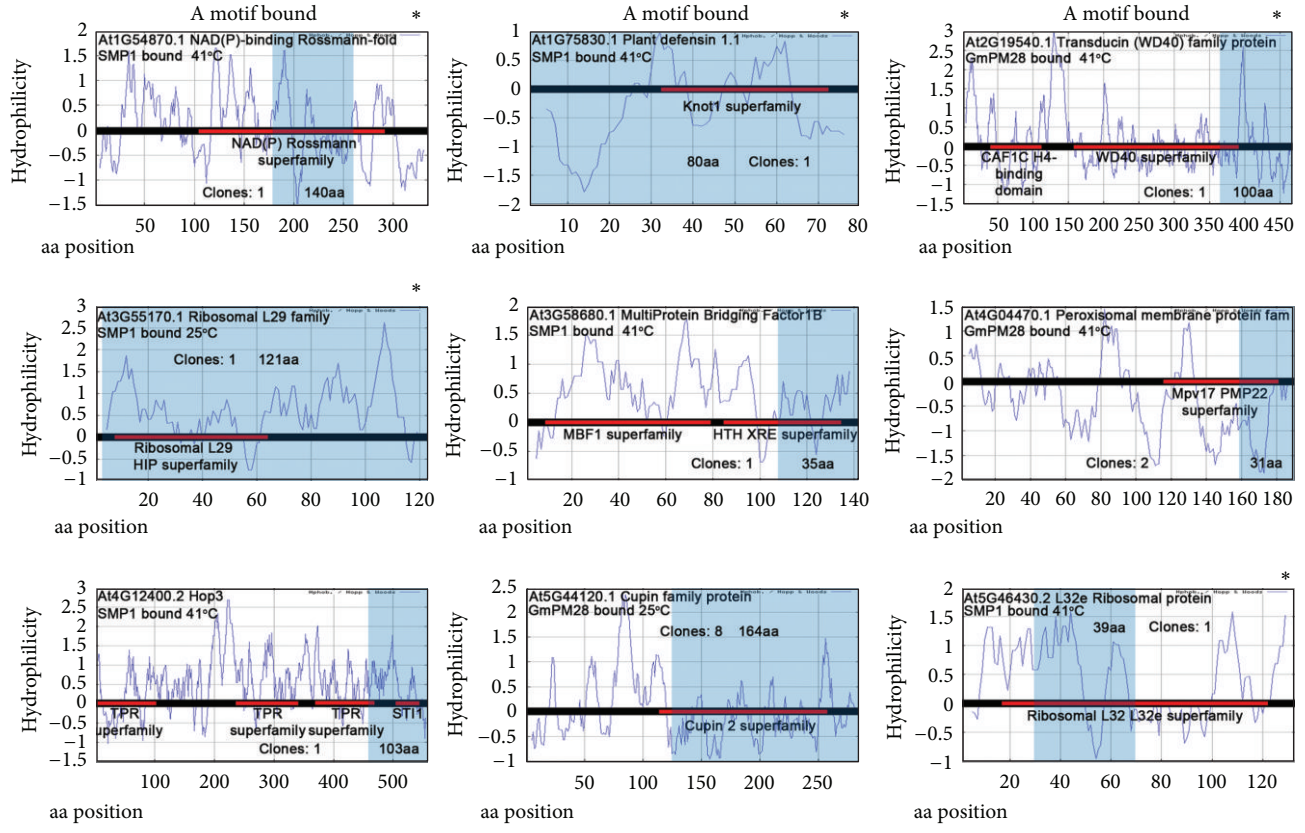
Model quality was determined using the Protein Structure and Model Assessment Tool available through the SWISS-MODEL server [32, 33]. The estimated absolute model quality is reported here using the QMEAN Z-score in Table 1, which is an estimate related to reference X-ray crystallographic structures [34]. The reported Z-scores are standard deviations of the homology model relative to expected values from experimental structures.

3. Results and Discussion

Determination of similarity within the LEA client protein subset begins with analysis of similarities both within the bound region and the full-length client protein. For each of the client proteins identified through phage display as described previously [25], Figure 1 illustrates the hydrophilicity profile (Hopp/Woods analysis from ProtScale [35]) of the entire protein along with any identifiable protein domains. The figure has been divided to categorize the client proteins into those binding the most hydrophilic regions, those in which an identifiable protein motif has been bound, and those with no recognizable attributes having been bound. This preliminary analysis and classification of LEA client proteins have overlapping category members (signified by an asterisk, Figure 1). From this analysis, it is not immediately

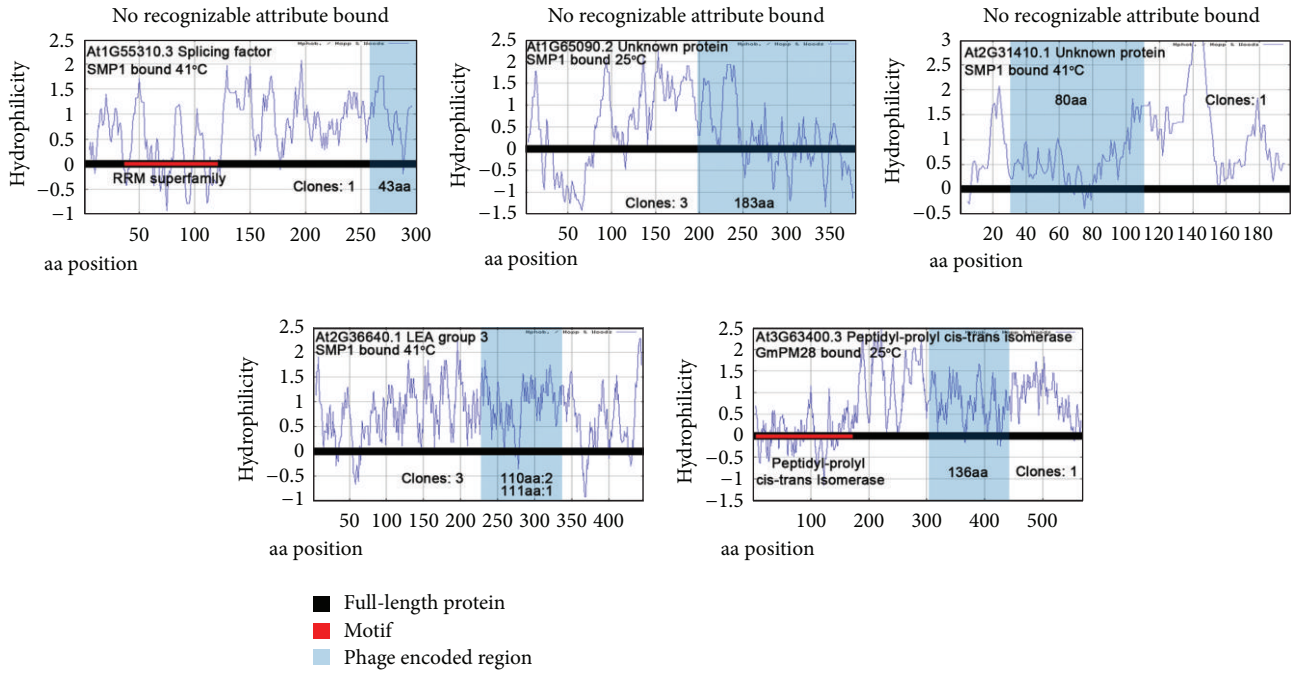


(a)



(b)

Figure 1: Continued.



(c)

Figure 1: A graphic depiction of the region of the client proteins to which the SMP1 or GmPM28 proteins bound. In each graph, the full-length protein is depicted as a black bar centered at zero on the Hopp/Woods hydrophilicity plot [57] for the protein (retrieved from Expasy Protscale [35]). Recognizable motifs present in the protein are represented as red bars on the black bar under which the Pfam [58] acronym defining the motif superfamily is displayed. The region of the full-length protein displayed on the phage and captured by the LEA is shaded grey. When this region overlaps with a recognizable motif, the protein is assigned to (b) (A motif bound). When it coincides with the most, or among the most, hydrophilic of the proteins regions, it is placed in (a) (or marked by an asterisk in (b)). If the LEA-bound fragment is neither the most hydrophilic nor encoding a recognizable motif, it is placed in (c) (no recognizable attribute bound). In each graph, the size, in amino acids, of the protein moiety bound by the LEA is provided as well as the number of independently acquired clones. If the clones were of different lengths, the number of clones of a specific length is provided. Whether the clone was bound by SMP1 or GmPM28 and the temperature at which the binding occurred are also provided.

clear what, if anything, this set of proteins has in common. The full-length proteins are wildly variable in length (79–1608 amino acids) as are the regions containing the portion of the proteins to which the LEAs bind (24–183 amino acids). This latter attribute is consistent with the use of random hexamers to synthesize the phage display libraries [36]. Furthermore, the identifiable protein motifs do not appear to have commonality, though several ribosomal proteins appeared to preferentially bind to the LEA proteins. Interpretation of the hydrophilicity profiles is also mystifying, because while often the most or among the most hydrophilic protein regions are recovered via phage display, exceptions exist.

Evaluation exclusive to the regions containing those bound by the LEA proteins provides more insight into what the LEA client proteins may have in common. Amino acid composition, normalized by length of the region (Figure 2(a)), reveals that the bound regions have relatively low occurrences of aromatic residues, Phe, Trp, Tyr, and His, and sulfur-containing residues, Cys and Met, lending a general hydrophilicity to the region. Such an attribute would be consistent with the solvent-exposed exterior of a globular protein to which the LEA protein is presumed to bind. Lack of surface-exposed, thiol-containing, amino acids is not surprising given their tendency towards oxidation. Interestingly, the amino acid composition profile is consistent with that of a protein-nucleic acid complex data set examined by Baker and Grant [37]. Baker and Grant postulate that despite the low prevalence of aromatic residues within the binding sites of protein-nucleic acid complexes, aromatic residues still play a critical role in nucleic acid recognition. Relative to our observations, however, the binding site amino acid frequency of protein-nucleic acid complexes appears to be dominated by Arg, Lys, Asn, Glu, Gly, Ser, Thr, and Asp residues, providing us with a common thread potentially linking this set of LEA client proteins.

Further analysis of the hydropathicity of the bound regions provides additional insight into potential functional relationships of the LEA client proteins. The grand average hydropathicity (GRAVY) was determined for each of the LEA client proteins as shown in Figure 2(b). The GRAVY hydropathicity is calculated based on the Kyte and Doolittle [38] hydropathy values for each amino acid, the total of which is subsequently divided by the number of amino acids in the sequence to arrive at an average [35]. A negative value indicates hydrophilicity, and likewise, a positive value indicates hydrophobicity. We see that for a vast majority of

Table 1: PDB templates used for each of the seven homology models of the LEA client proteins. The four-character PDB identifier is provided. The chain identifier follows the underscore. A brief description of each of the PDB template molecules is provided.

	PDB template	Description	Z-score
AT1G54870.1	3ijr_A (no publication)	*Bacillus anthracis* short chain dehydrogenase	−1.24
AT1G75830.1	1ayj_A [60]	*Raphanus sativu*s antifungal protein 1	−0.27
AT3G55170.1	2zkr_v [61]	Mammalian ribosomal 60S subunit	−0.64
	4a17_U [62]	*Tetrahymena thermophilia* 60S ribosomal subunit	
	3u5e_h [63]	Eukaryotic ribosome	
	3iz5_c [64]	*Triticum aestivum* ribosomal protein	
AT3G58680.1	3kxa_A [65]	*Neisseria gonorrhoeae* NGO0477	0.83
	2jvl_A [66]	*Trichoderma reesei* multiprotein bridging factor 1	
AT5G18380.1	3u5c_Q [63]	Eukaryotic ribosome	0.02
AT5G44120.1	3kgl_A [67]	*Brassica napus* 11S globulin, procruciferin	−2.32
AT5G46430.2	4a17_X [62]	*Tetrahymena thermophilia* 60S ribosomal subunit	−0.77

our LEA client proteins, the bound region is overwhelmingly hydrophilic, lending credence to the putative role of LEA proteins in the protection of client proteins from dehydration. However, two regions, part of AT1G75830.1 and AT4G04470.1, are identified as hydrophobic, and two others, AT1G65090.2 and AT5G44120.1, are only mildly hydrophilic. For all four of these bound regions, we confirmed that a single residue or subset of residues was not dominating the average, and rather, the hydrophobicity or mild hydrophilicity is indicative of the nature of the entire bound region (see Figure 1).

The entire set of full-length LEA client proteins was also analyzed using WOLF-pSORT (invoking the plant option), a program designed to predict protein localization sites [39]. Of the set, all but three were predicted to reside within a subcellular compartment containing nucleic acid polymers, with most predicted as either nuclear or cytoplasmic. The three outliers in the WOLF-pSORT analysis included AT1G75830.1, AT2G36640.1, and AT5G44120.1. AT1G75830.1 was predicted as extracellular. AT2G36640.1 was predicted to be peroxisomal, and AT5G44120.1 was predicted to be vacuolar. It is noteworthy that two of the three WOLF-pSORT outliers correspond to the hydrophobic or only mildly hydrophilic binding regions. This suggests, as does the amino acid composition, that perhaps an overall commonality of the remaining LEA client proteins is an ability to bind nucleic acids or at the very least promote interaction.

The amino acid sequences of the full-length LEA client proteins were analyzed using two separate protein motif/pattern and signature identification utilities with the aim of uncovering unifying motifs or functionality within the set. Table 2 summarizes the motifs and patterns uncovered, delineating between those that belong to the sequence region containing the bound moiety and those belonging to the full-length protein excluding this region. Putative amidation motifs (x-G-[RK]-[RK]) were identified using the patmat-motifs utility within the EMBOSS software suite, which searches amino acid sequences against the PROSITE motif database [40]. PROSITE defines an amidation site (PS00009) as situated at the carboxy terminus of an active peptide in a larger precursor protein at the site of cleavage. Typically, peptidylglycine α-amidating enzyme (α-AE) can utilize the amino group from the C-terminal glycine in this motif to effect the conversion of the amino acid "x" to an amidated-($CO–NH_2$) rather than a carboxylated-(COOH) terminus [41]. Nearly 60% of the full-length sequences contain at least one amidation domain (25% within the binding regions); however, relevance is difficult to determine at this point given the high natural probability of occurrence of this tetrapeptide sequence.

Using the InterProScan protein signature recognition software, potentially meaningful motifs, though not discernibly mutual, were established. The identification of the microbodies C-terminal targeting signal domain, a tripeptide C-terminal consensus sequence occasionally found in peroxisomal proteins [42], in AT2G36640.1 is consistent with the prediction from WOLF-pSORT of this protein as peroxisomal. This is not unexpected, as pSORT algorithms use the SKL motif as recognition mechanism for peroxisomal proteins. The RGD tripeptide sequence motif was also returned in three separate instances, which is thought to promote binding to integrins and similar proteins [43] and appears to be critical in mediation of cell attachment [44]. The leucine zipper and coiled coil motifs were also repeatedly returned by InterProScan searches. The leucine zipper is a protein-protein motif of α-helices that dimerizes to form a coiled coil. The leucine zipper is known to participate in DNA-binding and regulation of gene expression [45], and the coiled coil is suspected to more generally participate in protein-protein interactions [46]. Less often, though interesting, nonetheless, the ATP/GTP A motif was returned by InterProScan. This motif, ATP/GTP A, is a glycine-rich loop sequence connecting a β-strand and an α helix, which has been identified as a conserved region of ATP- and GTP-binding proteins through observation of crystallographic data [47–52]. The loop region is known to interact with the phosphate groups of nucleotides. Finally, several proteins were identified as ribosomal which, along with RNA in protein-RNA interactions, assemble to form ribosomal subunits [53]. While there does not seem to be a single unifying motif or pattern among the set of phage-display-identified LEA client proteins, there does appear to

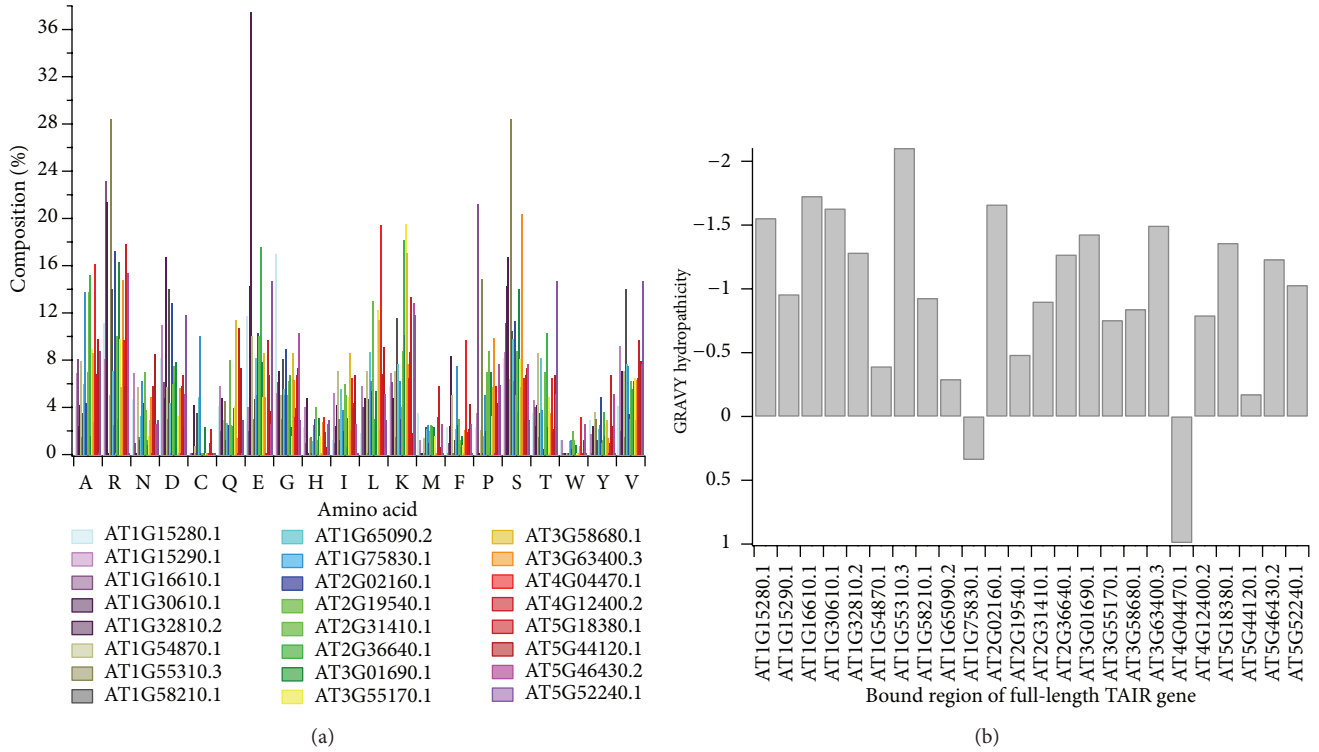


Figure 2: Analysis of the inclusive bound regions of the LEA client proteins identified using phage display. (a) Amino acid composition of the LEA-bound client protein regions is given here by % of the entire individual bound region. The regions are identified by the TAIR locus identifier for the full-length protein, though only composition of the bound region is represented in the plot. (b) A comparison of the GRAVY hydropathicity of the bound regions of the LEA client proteins is given here, again identified by the full-length protein TAIR locus identifier.

be a common thread of nucleotide interaction based upon known functionality of the identified patterns and motifs.

Within the amino acid sequences of the bound regions alone, PRATT was used to identify recurring patterns within the unaligned sequences (multiple sequence alignment of the diverse proteins being infeasible) [54]. Figure 3 illustrates the sequences of the bound LEA client protein regions, identified by the TAIR locus identifier. The sequences are coded with red and blue characters according to the two most commonly occurring patterns in the set of proteins. The K-x(2,4)-V-x(4)-[ACDGNSTV] pattern, represented in red, is found in 75% of the bound protein regions identified using phage display. In blue, the R-x(1,2)-R-x(0,1)-S pattern is common to 50% of the bound protein regions. PDBeMotif, a search algorithm providing statistics from 3D structural data, was used to interpret the significance of these two patterns [55]. For both patterns, PDBeMotif suggests—based on existing 3D structural data of proteins containing these sequence patterns—that glycerophosphate, ribose, and deoxyribose are among the structure-bound ligands. These sugars comprise the backbone of nucleic acid, and the presence of the phosphate in glycerophosphate is chemically consistent with an ester-linked phosphate of a nucleotide dimer (Figure 4). While this is by no means confirmation of the common functionality of the LEA client proteins, which can only be guaranteed through experimental means, sequence analysis and pattern/motif algorithms based on existing structural and functional data continually return to the theme of protein-nucleotide interactions.

The computational analysis of LEA client proteins concludes with homology modeling, where feasible, of the phage-display-bound regions of the LEA client proteins. As with the bioinformatics-based investigation, the intent here was to identify defining characteristics, either structural or chemical, which may yield insight as to why these proteins in particular are consistently returned as LEA-binding partners. Homology modeling methodology and identified structural templates are described in the methods section.. As alluded to the above, we were only able to successfully identify suitable 3D structural templates for seven of the LEA client proteins. Many of the LEA client proteins, including some of those modeled here, exist as membrane-bound proteins and are thus difficult to resolve structurally. The seven LEA client proteins include AT1G54870.1, AT1G75830.1, AT3G55170.1, AT3G58680.1, AT5G18380.1, AT5G44120.1, and AT5G46430.2. We anticipate that as crystallographic methods continue to develop, additional structures will become available to serve as templates to the remaining client proteins. Several other templates were available for portions of the full-length proteins; however, we are restricting this homology modeling study to that within the bound regions, as this should intuitively provide the most information

>AT1G15280.1
EVGTVKYDNDEDGEDSYEDDEEESGGGIDNDKSGVVKEAGDMNGEEENEKEKLQAAVPTG
GAFYMHDDRFQEMSAGGNRRMRGGRRQWGSGEERKWGHDKFEEMNTGEKHSDQRMSRGRF
RGHGRGRGQGRGYARGSSSNTLTSSGQQIYVPKAVSRGRGPRKSDTPLRNE

>AT1G15290.1
GIPKPDASIASKGHLSVSDLLDYISSDPDTKGNVAHRKHRRARILQVNDKVASADDDAHR
VASQIDIVTWNNVAEADVTKSRSEVNDPDTVVDKTNIETGDIVVHRLNVDRQTVEESTLD
EGWQEAYSKGRSGNGAGRKSRQRQPDLMKKRMLLNKHHNRNQDVQQQNIYSPL

>AT1G16610.1
FTLPPRQKVSSPPKPVSAAPKRDAPKSDNAAADAEKDGGPRRPRETSPQRKTGLSPRRRS
PLPRRGLSPRRRSPDSPHRRRPGSPIRRRGDTPPRRRPA

>AT1G30610.1
ESFRRRYSKQEHHRSSDTSRGIARGSKGDELELVVEERHVQR

>AT1G32810.2
EEEVSEEDEEDAFSDTSEESIFCD

>AT1G54870.1
TVVKGQEEKDAQETLQMLKEVKTSDSKEPIAIPTDLGFDENCKRVVDEVVNAFGRIDVLI
NNAAEQYESSTIEEIDEPRLERVFRTNIFSYFFLTRHALKHMKEGSSIINTTSVNAYKGN
ASLLDYTATKGAIVAFTRGL

>AT1G55310.3
ISRSPRRSRSPSPKRNRSVSPRRSISRSPRRSRSPRRSRRSYTPEPARSRSQSPHGGQYD
EDRSPSQ

>AT1G58210.1
VGSLRDLVCQKSDKACEKSRVGDVDDDDDDDDKSLVSRVVGVKTRGMLRRREEYEASIG
KRVATMSGKRVVTVSKKKNRRRSFGC

>AT1G65090.2
ATKIETSTGKDDEEISSNEPIDQASGAQGTGEEKRKNTTKKKKKTGRAGNRFLKCHTWSS
SKLCGRCDLLECCFDRVDCVVRRVITCSALSLISEASVSMSRICMVLNLQVYSEEQLWET
METLRKVVGYSVARSATCAEELKALYVFTGVVEPPRSSLNQDTYDIAHLTIRLRFLMSVI
GIN

>AT1G75830.1
MAKSATIVTLFFAALVFFAALEAPMVVEAQKLCERPSGTWSGVCGNSNACKNQCINLEKA
RHGSCNYVFPAHKCICYFPC

>AT2G02160.1
LQKYGSDDNNSFHNGKDADDVLRESSPGFDVLVDNEAGDSEYYHVEDRYGRRSQERGNSE
YDPDFSAIADGDKEALREQRFDSYDRREDRGWGHRRVSSEREDRLDRRVYAEDERSENIL
ESDLRYRLAKQRKGNGMRLSVGGHDYAAPDSSMDRGYRESRRDTPRENSISSSRLQGRIK
LRERSNGEEGHHFDRRSERGRDR

>AT2G19540.1
AHEASTLAVTSGDNQLTIWDLSLEKDEEEEAEFNAQTKELVNTPQDLPPQLLFVHQGQKD
LKELHWHNQIPGMIISTAGDGFNILMPYNIQNTLPSELPA

>AT2G31410.1
AIADEASMDIDGAPPAAKRSAVASSENPDKPIALAVERPTYDGVIAGKVSGRNWKQPRTH
RSSGRFVKNRKPDLEEMKRQ

>AT2G36640.1
EKAKETANYTADKAKEAKDKTAEKVGEYKDYTVDKAVEARDYTAEKAIEAKDKTAEKTGE
YKDYTVEKATEGKDVTVSKLGELKDSAVETAKRAMGFLSGKTEEAKGKAVETKDTAKENM
EKAGEVTRQKMEEMRLEGKELKEEAGAKAQEASQKTRESTESGAQ

>AT3G01690.1
PLWVKGGNHCDLEHYPEYIRHLKKFIATVERLPCPRMSSDQSERVRDAPPRRSMDRRVKP
RQSTERREKEKPPKSQSKMSSSSSKLKISFDQLDRSRRSVDCHEKTRKSVDQIERGRKSV
DRLDRVRSE

>AT3G55170.1
MARIKVHELRDKSKSDLSTQLKELKAELASLRVAKVTGGAPNKLSKIKVVRKSIAQVLTV
SSQKQKSALREAYKNKKLLPLDLRPKKTRAIRRRLTKHQASLKTEREKKKDMYFPIRKYA
IKV

>AT3G58680.1
KPQVIQEYESGKAIPNQQILSKLERALGAKLRGKK

>AT3G63400.3
SSDTESSSSSDDEKVGHKAIKSVKVDNADQHANLDDSVKSRSRSPIRRRNQNSRSKSPSR
SPVRVLGNGNRSPSRSPVRDLGNGSRSPREKPTEETVGKSFRSPSPSGVPKRIRKGRGFT
ERYSFARKYHTPSPERSPPRHW

>AT4G04470.1
RVILHSLVAFFWGIFLTLRARSMTLALAKAK

>AT4G12400.2
AMETYQEGLKHDPKNQEFLDGVRRCVEQINKASRGDLTPEELKERQAKAMQDPEVQNILS
DPVMRQVLVDFQENPKAAQEHMKNPMVMNKIQKLVSAGIVQVR

>AT5G18380.1
EQSKKEIKDILVRYDRTLLVADPRRCEPKKFGGRGARSRYQKSYR

>AT5G44120.1
QLGYISTLNSYDLPILRFIRLSALRGSIRQNAMVLPQWNANANAILYVTDGEAQIQIVND
NGNRVFDGQVSQGQLIAVPQGFSVVKRATSNRFQWVEFKTNANAQINTLAGRTSVLRGLP
LEVITNGFQISPEEARRVKFNTLETTLTHSSGPASYGRPRVAAA

>AT5G46430.2
ESWRRPKGIDSRVRRKFKGVTLMPNVGYGSDKKTRHYLP

>AT5G52240.1
ENVEQDAHVTTTPGKTVVDKSDDAPAETVLKKEE

Figure 3: Amino acid sequences of the bound regions of the LEA client proteins. Recurring patterns within the set of sequences have been identified by red and blue text. The red characters indicate the K-x(2,4)-V-x(4)-[ACDGNSTV] pattern. Blue characters indicate the R-x(1,2)-R-x(0,1)-S pattern.

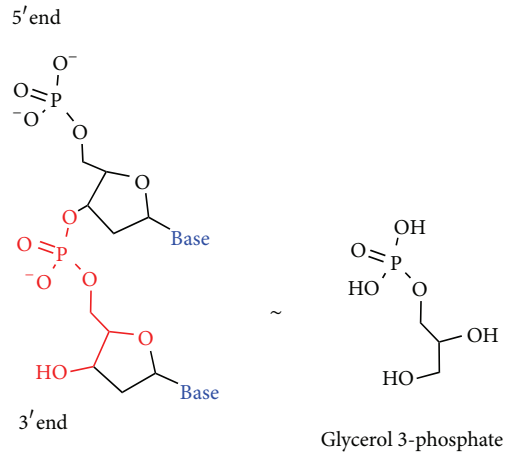


Figure 4: Chemical structure of a nucleotide dimer, left, and glycerol 3-phosphate (glycerophosphate), right. The red lettering on the nucleotide dimer represents the chemical similarity to the glycerophosphate molecule.

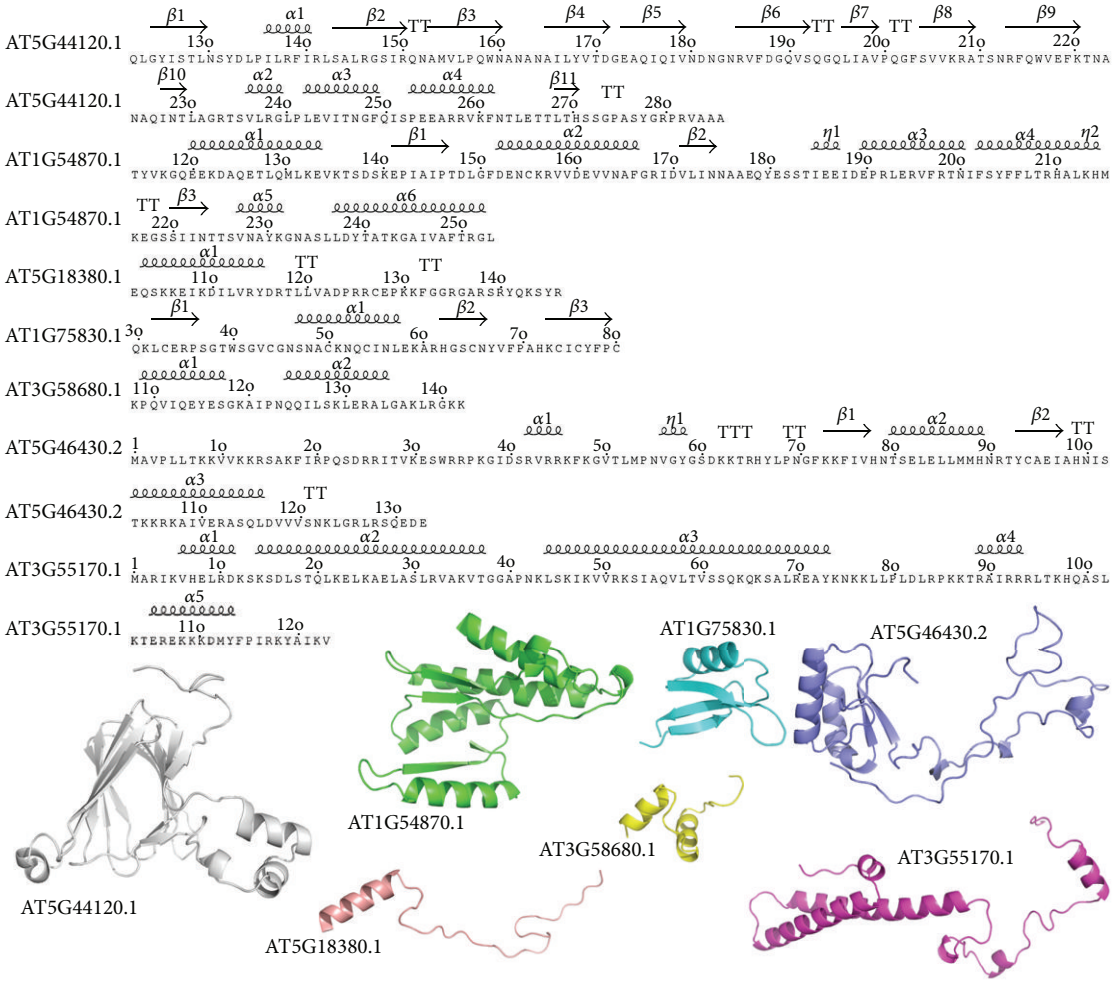


Figure 5: Seven homology models of LEA client proteins, focused on the regions containing the protein moiety to which the LEA proteins bound, were developed. The sequences, numbered by the full-length protein TAIR locus identifier, are shown annotated by secondary structure elements. Secondary structure annotation was accomplished using the ESPript web utility [59]. β-Sheets are labeled with a solid black arrow, α-helices with medium curly script, β-turns with TT, and 3_{10}-helices (η) with small curly script. Sequence number is also indicated in frequency of ten and corresponds to that of the full-length sequence. Below the sequences, the seven homology models of the bound regions only are shown in cartoon representation. The homology models are labeled, as with the sequences, according to the TAIR locus identifier of the full-length protein to which they belong. The homology model PDB files have been included in Supplementary Materials available online at http://dx.doi.org/10.1155/2013/470390.

regarding features contributing to the protein-protein interactions. Figure 5 illustrates the sequences of the seven homology models annotated according to the predicted secondary structure. Below the sequences, cartoon representations of each model are provided.

Structural or sequential representation of these seven protein regions does not provide a striking explanation as to which attribute is acting as a functional link. Several of the structures exhibit relatively large expanses of disordered loop regions, which seem rather uncharacteristic of globular proteins. This is almost certainly related to the hydrophilic nature of the bound regions (Figure 2(a)). We do find it somewhat intriguing, however, that of the three proteins returned by WOLF-pSORT as being neither nuclear nor cytoplasmic, two (AT3G55170.1 and AT1G75830.1) homologous protein structures are available through the Protein Data Bank, though we cannot ascribe significance to this based on the data presented here. With our limited subset of binding site homology models, we can only state that the structures appear to vary significantly from one another and that commonality may lie more in the chemical and dynamical rather than the structural nature of the region.

4. Conclusion

A great deal of information relating both sequence and structure of the LEA client protein bound regions and how this contributes to binding remains to be determined. From this initial computational study aiming to shed light on the functional role of LEA proteins through similarity in their bound substrates, we have uncovered what seems to be a predilection for protein-nucleic acid interaction in the

Table 2: Patterns and motifs identified using InterProScan and the patmatmotifs utility as part of the EMBOSS package. The full-length protein is identified by the TAIR locus identifier. Patterns and motifs have been separated according to their position either within the binding region or exclusive of the binding region.

	Full-length (excludes binding region)	Binding Region
AT1G15280.1	Amidation motif	Amidation motif
AT1G15290.1	Amidation motif (2) Leucine zipper Coiled coil	Amidation motif
AT1G16610.1	Amidation motif	RGD
AT1G30610.1	RGD ATP/GTP A	—
AT1G32810.2	Zinc finger plant homeodomain Amidation motif (4)	Coiled coil
AT1G54870.1	—	Short-chain dehydrogenase
AT1G55310.3	Amidation motif (2)	—
AT1G58210.1	Coiled coil (11) Leucine zipper	Amidation motif (2)
AT1G65090.2	Amidation motif	—
AT1G75830.1	—	Gamma thionin
AT2G02160.1	Amidation motif Coiled coil	Amidation motif
AT2G19540.1	Amidation motif	—
AT2G31410.1	Coiled coil	—
AT2G36640.1	Microbodies C-ter Coiled coil (2)	Coiled coil
AT3G01690.1	—	Amidation motif
AT3G55170.1	—	Ribosomal L29
AT3G58680.1	Coiled coil	Amidation motif
AT3G63400.3	Prolyl-peptidyl isomerase ATP/GTP A (2) Amidation motif (3)	—
AT4G04470.1	Amidation motif	—
AT4G12400.2	Coiled coil	RGD
AT5G18380.1	Amidation motif Ribosomal	—
AT5G44120.1	11s seed storage	—
AT5G46430.2	—	Ribosomal L32e
AT5G52240.1	—	—

LEA client proteins. While this does not yet tell us how the LEA proteins function relative to the bound protein regions, it does suggest hypotheses to be tested concerning the subcellular residence of the LEA proteins under study. An evolutionary relationship between the LEA protein and the substrate protein dictates that the SMP1 and GmPM28 homologs be located in subcellular compartments containing the nucleic acid polymers to which their client proteins apparently bind (i.e., the nucleus, cytoplasm, plastids, and/or mitochondria). Phenotypic consequences for specific LEA protein (or LEA protein family) reductions [36, 56], as well as a demonstration that LEA protein homologs from the Seed Maturation Protein family have preferred client proteins to which they bind [25], suggest that at least some LEA proteins are not redundantly backed up, indiscriminate spacer molecules, and lead to the conclusion that other LEA proteins will also have preferred binding partners. The elucidation of the subfunctionalization of specific LEA proteins concerning which client proteins they bind to is most efficaciously performed using phage display.

In the near term, molecular dynamics simulations of the LEA client protein homology models, including the full-length domains, may provide additional insight into the flexibility and solvation dynamics of the proteins, in addition to directing ongoing experimental phage display efforts. The long-term focus will be on the development of additional homology models as more crystallographic structures become available as well as *de novo* protein design using rapidly developing structure prediction methods. Our continuing aim is the effective integration of computational modeling with phage display for the prediction of protein structures at risk for dehydration or heat damage, uncovering the mechanisms by which LEA proteins perform their protective function. Future endeavors could conceivably encompass phage-display-directed evolution of synthetic LEA proteins engineered to protect labile proteins.

Acknowledgments

This project was partially funded by an NSF IOS (0849230), Hatch, McIntire-Stennis (AD421 CRIS), USDA Seed Grant (2011-04375), and Sir Frederick McMaster Research Fellowship to A. Bruce Downie. The authors thank Stephen Chmely for his assistance with figure preparation.

References

[1] J. S. Clegg, "Cryptobiosis—a peculiar state of biological organization," *Comparative Biochemistry and Physiology*, vol. 128, no. 4, pp. 613–624, 2001.

[2] J. H. Crowe, J. F. Carpenter, and L. M. Crowe, "The role of vitrification in anhydrobiosis," *Annual Review of Physiology*, vol. 60, pp. 73–103, 1998.

[3] J. Buitink, O. Leprince, M. A. Hemminga, and F. A. Hoekstra, "Molecular mobility in the cytoplasm: an approach to describe and predict lifespan of dry germplasm," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 5, pp. 2385–2390, 2000.

[4] W. Q. Sun and A. C. Leopold, "Cytoplasmic vitrification and survival of anhydrobiotic organisms," *Comparative Biochemistry and Physiology A*, vol. 117, no. 3, pp. 327–333, 1997.

[5] O. Leprince, F. J. M. Harren, J. Buitink, M. Alberda, and F. A. Hoekstra, "Metabolic dysfunction and unabated respiration precede the loss of membrane integrity during dehydration of germinating radicles," *Plant Physiology*, vol. 122, no. 2, pp. 597–608, 2000.

[6] J. Buitink and O. Leprince, "Glass formation in plant anhydrobiotes: survival in the dry state," *Cryobiology*, vol. 48, no. 3, pp. 215–228, 2004.

[7] W. F. Wolkers, S. McCready, W. F. Brandt, G. G. Lindsey, and F. A. Hoekstra, "Isolation and characterization of a D-7 LEA

protein from pollen that stabilizes glasses in vitro," *Biochimica et Biophysica Acta*, vol. 1544, no. 1-2, pp. 196–206, 2001.

[8] K. Goyal, L. J. Walton, and A. Tunnacliffe, "LEA proteins prevent protein aggregation due to water stress," *Biochemical Journal*, vol. 388, part 1, pp. 151–157, 2005.

[9] V. Boucher, J. Buitink, X. Lin et al., "MtPM25 is an atypical hydrophobic late embryogenesis-abundant protein that dissociates cold and desiccation-aggregated proteins," *Plant, Cell and Environment*, vol. 33, no. 3, pp. 418–430, 2010.

[10] L. Dure III, S. C. Greenway, and G. A. Galau, "Developmental biochemistry of cottonseed embryogenesis and germination: changing messenger ribonucleic acid populations as shown by in vitro and in vivo protein synthesis," *Biochemistry*, vol. 20, no. 14, pp. 4162–4168, 1981.

[11] G. A. Galau, D. W. Hughes, and L. I. Dure, "Developmental biochemistry of cottonseed embryogenesis and germination: changing messenger ribonucleic acid populations as shown by reciprocal heterologous complementary deoxyribonucleic acid-messenger ribonucleic acid hybridizationembryogenesis-abundant (LEA) mRNAs," *Plant Molecular Biology*, vol. 7, pp. 155–170, 1986.

[12] J. M. Mouillon, P. Gustafsson, and P. Harryson, "Structural investigation of disordered stress proteins. Comparison of full-length dehydrins with isolated peptides of their conserved segments," *Plant Physiology*, vol. 141, no. 2, pp. 638–650, 2006.

[13] S. C. Hand, M. A. Menze, M. Toner, L. Boswell, and D. Moore, "LEA proteins during water stress: not just for plants anymore," *Annual Review of Physiology*, vol. 73, pp. 115–134, 2011.

[14] S. Singh, C. C. Cornilescu, R. C. Tyler et al., "Solution structure of a late embryogenesis abundant protein (LEA14) from *Arabidopsis thaliana*, a cellular stress-related protein," *Protein Science*, vol. 14, no. 10, pp. 2601–2609, 2005.

[15] D. Tolleter, M. Jaquinod, C. Mangavel et al., "Structure and function of a mitochondrial late embryogenesis abundant protein are revealed by desiccation," *Plant Cell*, vol. 19, no. 5, pp. 1580–1589, 2007.

[16] J. Eom, W. R. Baker, A. Kintanar, and E. S. Wurtele, "The embryo-specific EMB-1 protein of *Daucus carota* is flexible and unstructured in solution," *Plant Science*, vol. 115, no. 1, pp. 17–24, 1996.

[17] J. L. Soulages, K. Kim, E. L. Arrese, C. Walters, and J. C. Cushman, "Conformation of a group 2 late embryogenesis abundant protein from soybean. Evidence of poly (L-proline)-type II structure," *Plant Physiology*, vol. 131, no. 3, pp. 963–975, 2003.

[18] J. L. Soulages, K. Kim, C. Walters, and J. C. Cushman, "Temperature-induced extended helix/random coil transitions in a group 1 late embryogenesis-abundant protein from soybean," *Plant Physiology*, vol. 128, no. 3, pp. 822–832, 2002.

[19] T. Lisse, D. Bartels, H. R. Kalbitzer, and R. Jaenicke, "The recombinant dehydrin-like desiccation stress protein from the resurrection plant *Craterostigma plantagineum* displays no defined three-dimensional structure in its native state," *Biological Chemistry*, vol. 377, no. 9, pp. 555–561, 1996.

[20] P. S. Russouw, J. Farrant, W. Brandt, and G. G. Lindsey, "The most prevalent protein in a heat-treated extract of pea (*Pisum sativum*) embryos is an LEA group I protein; its conformation is not affected by exposure to high temperature," *Seed Science Research*, vol. 7, no. 2, pp. 117–123, 1997.

[21] A. M. Ismail, A. E. Hall, and T. J. Close, "Purification and partial characterization of a dehydrin involved in chilling tolerance during seedling emergence of cowpea," *Plant Physiology*, vol. 120, no. 1, pp. 237–244, 1999.

[22] C. B. F. Andersen, L. Ballut, J. S. Johansen et al., "Structure of the exon junction core complex with a trapped DEAD-Box ATPase bound to RNA," *Science*, vol. 313, no. 5795, pp. 1968–1972, 2006.

[23] M. J. Howard, W. H. Lim, C. A. Fierke, and M. Koutmos, "Mitochondrial ribonuclease P structure provides insight into the evolution of catalytic strategies for precursor-tRNA 5' processing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 40, pp. 16149–16154, 2012.

[24] F. Kiefer, K. Arnold, M. Künzli, L. Bordoli, and T. Schwede, "The SWISS-MODEL repository and associated resources," *Nucleic Acids Research*, vol. 37, no. 1, pp. D387–D392, 2009.

[25] R. Kushwaha, T. D. Lloyd, K. R. Schäfermeyer, S. Kumar, and A. B. Downie, "Identification of late embryogenesis abundant (LEA) protein putative interactors using phage display," *International Journal of Molecular Sciences*, vol. 13, no. 6, pp. 6582–6603, 2012.

[26] A. Biegert, C. Mayer, M. Remmert, J. Söding, and A. N. Lupas, "The MPI Bioinformatics Toolkit for protein sequence analysis," *Nucleic Acids Research*, vol. 34, pp. W335–W339, 2006.

[27] J. Söding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2005.

[28] J. Söding, A. Biegert, and A. N. Lupas, "The HHpred interactive server for protein homology detection and structure prediction," *Nucleic Acids Research*, vol. 33, no. 2, pp. W244–W248, 2005.

[29] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature Methods*, vol. 9, no. 2, pp. 173–175, 2012.

[30] A. Šali, "Comparative protein modeling by satisfaction of spatial restraints," *Molecular Medicine Today*, vol. 1, no. 6, pp. 270–277, 1995.

[31] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[32] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling," *Bioinformatics*, vol. 22, no. 2, pp. 195–201, 2006.

[33] L. Bordoli, F. Kiefer, K. Arnold, P. Benkert, J. Battey, and T. Schwede, "Protein structure homology modeling using SWISS-MODEL workspace," *Nature Protocols*, vol. 4, no. 1, pp. 1–13, 2009.

[34] P. Benkert, M. Biasini, and T. Schwede, "Toward the estimation of the absolute quality of individual protein structure models," *Bioinformatics*, vol. 27, no. 3, pp. 343–350, 2011.

[35] E. Gasteiger, C. Hoogland, A. Gattiker et al., "Protein identification and analysis tools on the ExPASy server," in *The Proteomics Protocols Handbook*, J. M. Walker, Ed., pp. 571–607, Humana Press, New Jersey, NJ, USA, 2005.

[36] T. Chen, N. Nayak, S. M. Majee et al., "Substrates of the *Arabidopsis thaliana* protein isoaspartyl methyltransferase 1 identified using phage display and biopanning," *The Journal of Biological Chemistry*, vol. 285, no. 48, pp. 37281–37292, 2010.

[37] C. M. Baker and G. H. Grant, "Role of aromatic amino acids in protein-nucleic acid recognition," *Biopolymers*, vol. 85, no. 5-6, pp. 456–470, 2007.

[38] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, 1982.

[39] P. Horton, K. J. Park, T. Obayashi et al., "WoLF PSORT: protein localization predictor," *Nucleic Acids Research*, vol. 35, pp. W585–587, 2007.

[40] P. Rice, L. Longden, and A. Bleasby, "EMBOSS: the European molecular biology open software suite," *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, 2000.

[41] D. J. Merkler, "C-terminal amidated peptides: production by the in vitro enzymatic amidation of glycine-extended peptides and the importance of the amide to bioactivity," *Enzyme and Microbial Technology*, vol. 16, no. 6, pp. 450–456, 1994.

[42] S. J. Gould, G. A. Keller, and S. Subramani, "Identification of peroxisomal targeting signals located at the carboxy terminus of four peroxisomal proteins," *Journal of Cell Biology*, vol. 107, no. 3, pp. 897–905, 1988.

[43] G. B. Monshausen and S. Gilroy, "Feeling green: mechanosensing in plants," *Trends in Cell Biology*, vol. 19, no. 5, pp. 228–235, 2009.

[44] S. E. D'Souza, M. H. Ginsberg, and E. F. Plow, "Arginyl-glycyl-aspartic acid (RGD): a cell adhesion motif," *Trends in Biochemical Sciences*, vol. 16, no. 7, pp. 246–250, 1991.

[45] D. Krylov and C. R. Vinson, "Leucine zipper," in *Els*, pp. 1–7, John Wiley & Sons, New York, NY, USA, 2001.

[46] A. Singh and S. E. Hitchcock-Degregori, "Dual requirement for flexibility and specificity for binding of the coiled-coil tropomyosin to its target, actin," *Structure*, vol. 14, no. 1, pp. 43–50, 2006.

[47] T. E. Dever, M. J. Glynias, and W. C. Merrick, "GTP-binding domain: three consensus sequence elements with distinct spacing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 7, pp. 1814–1818, 1987.

[48] D. C. Fry, S. A. Kuby, and A. S. Mildvan, "ATP-binding site of adenylate kinase: mechanistic implications of its homology with ras-encoded p21, F1-ATPase, and other nucleotide-binding proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 83, no. 4, pp. 907–911, 1986.

[49] E. V. Koonin, "A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif," *Journal of Molecular Biology*, vol. 229, no. 4, pp. 1165–1174, 1993.

[50] W. Moller and R. Amons, "Phosphate-binding sequences in nucleotide-binding proteins," *FEBS Letters*, vol. 186, no. 1, pp. 1–7, 1985.

[51] M. Saraste, P. R. Sibbald, and A. Wittinghofer, "The P-loop—a common motif in ATP- and GTP-binding proteins," *Trends in Biochemical Sciences*, vol. 15, no. 11, pp. 430–434, 1990.

[52] J. E. Walker, M. Saraste, M. J. Runswick, and N. J. Gay, "Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold," *The EMBO Journal*, vol. 1, no. 8, pp. 945–951, 1982.

[53] D. J. Klein, P. B. Moore, and T. A. Steitz, "The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit," *Journal of Molecular Biology*, vol. 340, no. 1, pp. 141–177, 2004.

[54] I. Jonassen, J. F. Collins, and D. G. Higgins, "Finding flexible patterns in unaligned protein sequences," *Protein Science*, vol. 4, no. 8, pp. 1587–1595, 1995.

[55] A. Golovin and K. Henrick, "MSDmotif: exploring protein sites and motifs," *BMC Bioinformatics*, vol. 9, article 312, 2008.

[56] Y. Olvera-Carrillo, F. Campos, J. L. Reyes, A. Garciarrubio, and A. A. Covarrubias, "Functional analysis of the group 4 late embryogenesis abundant proteins reveals their relevance in the adaptive response during water deficit in arabidopsis," *Plant Physiology*, vol. 154, no. 1, pp. 373–390, 2010.

[57] T. P. Hopp and K. R. Woods, "Prediction of protein antigenic determinants from amino acid sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 78, no. 6 I, pp. 3824–3828, 1981.

[58] R. D. Finn, J. Mistry, J. Tate et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 38, no. 1, pp. D211–D222, 2010.

[59] P. Gouet, E. Courcelle, D. I. Stuart, and F. Métoz, "ESPript: analysis of multiple sequence alignments in PostScript," *Bioinformatics*, vol. 15, no. 4, pp. 305–308, 1999.

[60] F. Fant, W. Vranken, W. Broekaert, and F. Borremans, "Determination of the three-dimensional solution structure of *Raphanus sativus* antifungal protein 1 by ^{1}H NMR," *Journal of Molecular Biology*, vol. 279, no. 1, pp. 257–270, 1998.

[61] P. Chandramouli, M. Topf, J. F. Ménétret et al., "Structure of the mammalian 80S ribosome at 8.7 Å resolution," *Structure*, vol. 16, no. 4, pp. 535–548, 2008.

[62] S. Klinge, F. Voigts-Hoffmann, M. Leibundgut, S. Arpagaus, and N. Ban, "Crystal structure of the eukaryotic 60S ribosomal subunit in complex with initiation factor 6," *Science*, vol. 334, no. 6058, pp. 941–948, 2011.

[63] A. Ben-Shem, N. G. De Loubresse, S. Melnikov, L. Jenner, G. Yusupova, and M. Yusupov, "The structure of the eukaryotic ribosome at 3.0 Å resolution," *Science*, vol. 334, no. 6062, pp. 1524–1529, 2011.

[64] J. P. Armache, A. Jarasch, A. M. Anger et al., "Localization of eukaryote-specific ribosomal proteins in a 5. 5-Å cryo-EM map of the 80S eukaryotic ribosome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 46, pp. 19754–19759, 2010.

[65] J. Ren, S. Samshury, J. E. Nettleship, N. J. Saunders, and R. J. Owens, "The crytal structure of NGOO47 from Neisseria gonorrhoeae reveals a novel protein fold incorporating a helix-turn-helix motif," *Proteins*, vol. 78, no. 7, pp. 1798–1802, 2010.

[66] R. K. Salinas, C. M. Camilo, S. Tomaselli et al., "Solution structure of the C-terminal domain of multiprotein bridging factor 1 (MBF1) of *Trichoderma reesei*," *Proteins*, vol. 75, no. 2, pp. 518–523, 2009.

[67] M. R. G. Tandang-Silvas, T. Fukuda, C. Fukuda et al., "Conservation and divergence on plant seed 11S globulins based on crystal structures," *Biochimica et Biophysica Acta*, vol. 1804, no. 7, pp. 1432–1442, 2010.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 530696, 6 pages
http://dx.doi.org/10.1155/2013/530696

Research Article

Naïve Bayes Classifier with Feature Selection to Identify Phage Virion Proteins

Peng-Mian Feng,[1] Hui Ding,[2] Wei Chen,[3] and Hao Lin[2]

[1] *School of Public Health, Hebei United University, Tangshan 063000, China*
[2] *Key Laboratory for Neuroinformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China*
[3] *Department of Physics, School of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China*

Correspondence should be addressed to Wei Chen; greatchen@heuu.edu.cn and Hao Lin; hlin@uestc.edu.cn

Received 10 March 2013; Revised 16 April 2013; Accepted 28 April 2013

Academic Editor: Yanxin Huang

Knowledge about the protein composition of phage virions is a key step to understand the functions of phage virion proteins. However, the experimental method to identify virion proteins is time consuming and expensive. Thus, it is highly desirable to develop novel computational methods for phage virion protein identification. In this study, a Naïve Bayes based method was proposed to predict phage virion proteins using amino acid composition and dipeptide composition. In order to remove redundant information, a novel feature selection technique was employed to single out optimized features. In the jackknife test, the proposed method achieved an accuracy of 79.15% for phage virion and nonvirion proteins classification, which are superior to that of other state-of-the-art classifiers. These results indicate that the proposed method could be as an effective and promising high-throughput method in phage proteomics research.

1. Introduction

Phage is a virus that infects and replicates within bacteria. Phages are widely distributed in locations populated by bacterial hosts, such as soil or the intestines of animals. A complete infectious phage viral particle (also, namely, phage virion) consists of an inner core of nucleic acid which gives the virus infectivity and a protein coat (called a capsid) which encases the nucleic acid and provides specificity, that is, determines which organisms the virus can infect.

The nucleic acid of phage virions is either RNA or DNA. Proteins of phage virions include structural proteins and nonstructural proteins. Structural proteins commonly termed "phage virion proteins" are essential materials of the infectious viral particles, including shell proteins, envelope proteins, and virus particle enzymes. Nonstructural proteins (namely, phage nonvirion proteins) refer to that encoded by the viral genome and play important roles in biological process of viral genome replication and expression, but they do not bind to phage virions. Due to the distinct functions between phage virion proteins and phage nonvirion proteins, knowledge about the protein composition of phage virions is an essential step to further understand the functions of phage virions.

Although the use of mass spectrometry (MS) for the identification of phage virion proteins has become popular [1], it has not kept pace with the explosive growth of protein sequences generated in the postgenomic age. Hence, it is highly desired to develop automated methods for timely and reliably classifying the protein composition of phage virions.

To the best of our knowledge, there is no computational system for the classification of phage virion proteins. In the current study, we propose a Naïve Bayes based computational model for predicting phage virion proteins using amino acid compositions and dipeptide compositions. The correlation-based feature subset selection algorithm [2] was introduced to find the optimal feature set. By using the optimized features, the proposed model was evaluated in a benchmark dataset

in the jackknife test. The performance demonstrates that this model could be a potentially useful tool for the annotation of the phage proteins.

According to some recent comprehensive reviews [3, 4] and demonstrated by a series of recent publications [5–10], to establish a really useful statistical predictor, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the statistical samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web server for the predictor that is accessible to the public. In the following, let us describe how to deal with these steps one by one.

2. Materials and Methods

2.1. Dataset. The raw datasets adopted in this research were extracted from the UniProt [11]. For the purpose of obtaining a reliable benchmark dataset, the following steps were considered. Firstly, only the experimentally confirmed phage virion and phage nonvirion protein sequences were included. Secondly, the sequences which are fragments of other proteins were dislodged. Thirdly, sequences containing nonstandard letters, that is, "B," "X," or "Z," were excluded as their meanings are ambiguous. After following the previous strict screening procedures, we obtained 121 phage virion protein sequences and 231 phage nonvirion protein sequences.

To prepare a high quality dataset, the CD-HIT program [12] was used to prune the data. By setting the cutoff of sequence identity to 40%, 307 sequences were remained in the final benchmark dataset, including 99 phage virion protein sequences and 208 phage nonvirion protein sequences.

2.2. Feature Vector. One of the most important parts for identifying protein attributes is to generate a set of proper informative parameters to encode the protein sequences. To avoid completely losing the sequence-order information, the pseudo amino acid composition (PseAAC) was proposed [13, 14] to replace the simple amino acid composition (AAC) for representing the sample of a protein. Since the concept of PseAAC was proposed in 2001 [13], it has been widely used to study various attributes of proteins, such as identifying bacterial virulent proteins [15], predicting supersecondary structure [16], predicting protein subcellular location [16–19], predicting membrane protein types [20], discriminating outer membrane proteins [21], identifying antibacterial peptides [22], identifying allergenic proteins [23], predicting metalloproteinase family [24], predicting protein structural class [25], identifying GPCRs and their types [26], identifying protein quaternary structural attributes [27], predicting protein submitochondria locations [28], identifying risk type of human papillomaviruses [29], identifying cyclin proteins [30], predicting GABA(A) receptor proteins [31], and classifying amino acids [32], among many others (see a long list of papers cited in the References section of [3]). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [7, 9], as well as other biological samples (see, e.g., [33, 34]). Because it has been widely and increasingly used, recently two powerful softwares, called "PseAAC-Builder" [35] and "propy" [36], were established for generating various special Chou's pseudoamino acid compositions.

The amino acid composition and dipeptide composition are the general forms of PseAAC and the simplest parameters, which also have been widely applied in the realm of protein prediction [37–40]. Hence, every protein sequence in the benchmark dataset was encoded in a discrete vector as

$$\mathbf{F} = [f_1, f_2, \ldots, f_{420}]^T, \quad (1)$$

where f_i is the normalized occurrence frequencies of the 20 amino acids ($i = 1, 2, \ldots, 20$) and the 400 dipeptides ($i = 21, 22, \ldots, 420$) in the protein sequence, respectively. T is the transposing operator.

2.3. Feature Selection. Inclusion of redundant and noisy features in the model building process would cause poor predictive performance and increased computation. Feature selection is the process of removing irrelevant features and is extremely useful in reducing the dimensionality of the data and improving the predictive accuracy. To reduce the dimension of the feature space and improve the precision of phage virion and nonvirion protein classification, the filter method Correlation-based Feature Selection [2] combined with Best-first search strategy was used in the process of feature selection in the current work.

The process starts with an empty set of features and generates all possible single feature expansions. The subset with the highest accuracy is chosen and expanded in the same way by adding single features. If the accuracy does not maximize with the expansion of a subset, the search drops back to the next best unexpanded subset and continues from there until all features are added. The subset with the highest accuracy will be selected as the final optimized feature set [41].

2.4. Naïve Bayes. Naïve Bayes is an effective statistical classification algorithm [42] and has been successfully used in the realm of bioinformatics [43–46]. The basic theory of Naïve Bayes is similar to that of Covariance Determinant (CD) [47–52]. But for Naïve Bayes, it assumes the attribute variables to be independent from each other given the outcome. This assumption greatly simplifies the calculation of conditional probabilities and also overcomes the divergent problem when using the CD prediction engine to deal with those systems in which the components of constituent feature vectors are normalized.

In the Naïve Bayes framework, a classification problem can be seen as the problem of finding the outcome with maximum probability given a set of observed variables. Given a phage viral protein example, described by its feature vector $\mathbf{F} = (f_1, f_2, \ldots, f_n)$, we are looking for a class $\mathbf{C}$ that maximizes the likelihood $\mathbf{P}(\mathbf{F} \mid \mathbf{C}) = \mathbf{P}(f_1, f_2, \ldots, f_n \mid \mathbf{C})$.

Since the current work is intend to classify phage virion and nonvirion proteins, a binary class $\mathbf{C} \in \{0, 1\}$ was generated, where 1 denotes that the sample was predicted as a phage virion protein and 0 denotes phage nonvirion protein. For the binary classification, the class for the protein sample could be determined by comparing two posteriors as

$$\frac{P\left(C = 1 \mid F = f_1, f_2, \ldots, f_n\right)}{P\left(C = 0 \mid F = f_1, f_2, \ldots, f_n\right)} = \frac{P\left(C = 1\right) \prod_{i=1}^{n} P_i\left(f_i \mid C = 1\right)}{P\left(C = 0\right) \prod_{i=1}^{n} P_i\left(f_i \mid C = 0\right)}. \tag{2}$$

Taking the logarithm of (2), we obtain

$$\log \frac{P\left(C = 1 \mid F = f_1, f_2, \ldots, f_n\right)}{P\left(C = 0 \mid F = f_1, f_2, \ldots, f_n\right)} = \log \frac{P\left(C = 1\right)}{P\left(C = 0\right)} + \sum_{i=1}^{n} \log \frac{P_i\left(f_i \mid C = 1\right)}{P_i\left(f_i \mid C = 0\right)}. \tag{3}$$

Hence the sample will be predicted as 1 (phage virion protein) if

$$\log \frac{P\left(C = 1 \mid F = f_1, f_2, \ldots, f_n\right)}{P\left(C = 0 \mid F = f_1, f_2, \ldots, f_n\right)} \geq \theta \tag{4}$$

and 0 (phage nonvirion protein) for otherwise. θ is the threshold determining the trade-off between sensitivity and specificity and can be trained on the training dataset to maximize the prediction performance.

2.5. Performance Evaluation. The performance of the proposed model was evaluated using sensitivity (Sn), specificity (Sp), and accuracy (Acc), which are expressed as

$$\begin{aligned} \text{Sn} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Sp} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{Acc} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}. \end{aligned} \tag{5}$$

TP, TN, FP, and FN represent the number of the correctly recognized phage virion proteins, the number of the correctly recognized phage nonvirion proteins, the number of phage nonvirion proteins recognized as phage virion proteins, and the number of phage virion proteins recognized as phage nonvirion proteins, respectively.

As the performance of the current classifier depends on the threshold θ as given in (4), the threshold independent parameter, receiver operating characteristic curve, was employed as well. Therefore, the quality of a classifier can be objectively evaluated by measuring the area under the receiver operating characteristic curve (auROC). The value of auROC score ranges from 0 to 1, with a score of 0.5 corresponding to a random guess and a score of 1.0 indicating a perfect separation.

Table 1: Predictive performance of Naïve Bayes based on different features.

Feature dimensions	Sn (%)	Sp (%)	Acc (%)	auROC
420	53.54	83.17	75.57	0.758
38	75.76	80.77	79.15	0.855

3. Results and Discussion

Three cross-validation methods, namely, subsampling test, independent dataset test, and jackknife test, are often employed to evaluate the predictive capability of a predictor. Among the three methods, the jackknife test is deemed the most objective and rigorous one that can always yield a unique outcome as demonstrated by a penetrating analysis in a recent comprehensive review [53] and hence has been widely and increasingly adopted by investigators to examine the quality of various predictors (see, e.g., [7, 19, 21, 30, 54–56]). Accordingly, the jackknife test was used to examine the performance of the model proposed in the current study. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent test sample, and all the rule parameters are calculated without including the one being identified.

3.1. Prediction of Phage Virion Proteins. We trained the Naïve Bayes classifier using Waikato Environment for Knowledge Analysis (WEKA) [57] on the benchmark dataset. As shown in Table 1, an auROC score of 0.758 and an accuracy of 75.57% with an average sensitivity of 53.54% and an average specificity of 83.17% were obtained for the classification of phage virion and nonvirion proteins by using all the 420 features, that is, 20 amino acid compositions and 400 dipeptide compositions.

In order to identify prominent features that can distinguish between phage virion and nonvirion proteins, feature selection method as introduced in Section 2.3 was carried out to eliminate the redundant features using WEKA in a tenfold cross-validation approach on the benchmark dataset. We found that the proposed method achieved a maximum accuracy of 79.48% and auROC of 0.86 when the feature dimension reduced to 38 (i.e., V, T, A, H, K, E, R, S, LE, VT, VG, MK, TA, TS, AT, HI, KL, KI, KH, KN, KK, KD, KE, KW, KR, DK, EF, EL, EV, EK, EE, EW, CE, WK, RE, SG, GV, and GG). The jackknife test results of the Naïve Bayes classifier based on the 38 optimized features were listed in Table 1. As it can be seen from Table 1, the current method yielded a best auROC score of 0.855 and a predictive accuracy of 79.15% with an average sensitivity of 75.76% and an average specificity of 80.77% (Table 1). Both predictive accuracy and auROC are higher than that of the model based on the 420 features.

3.2. Comparison with Other Methods. To the best of our knowledge, there exists no theoretical method for phage virion and nonvirion protein classifications. Therefore, we cannot provide the comparison analysis with published results to confirm that the model proposed here is superior to

Table 2: Comparison of Naïve Bayes with other methods by using optimized features.

Classifier	Sn (%)	Sp (%)	Acc (%)	auROC
BayesNet	68.69	79.81	76.22	0.799
RBFnetwork	72.73	82.21	79.15	0.839
Random Forest	55.56	84.62	75.24	0.802
LogitBoot	52.53	85.10	74.59	0.795
SVM	63.64	86.54	79.15	0.836
J48	61.62	77.88	72.64	0.671
Naïve Bayes	75.76	80.77	79.15	0.855

other methods. However, the proposed Naïve Bayes classifier was compared with other state-of-the-art classifiers, that is, BayesNet, RBFnetwork, Random Forest, J48, Support Vector Machine (SVM), and LogitBoot. All the classifiers were compared on the benchmark dataset based on the optimized features (i.e., V, T, A, H, K, E, R, S, LE, VT, VG, MK, TA, TS, AT, HI, KL, KI, KH, KN, KK, KD, KE, KW, KR, DK, EF, EL, EV, EK, EE, EW, CE, WK, RE, SG, GV, and GG). Their best predictive results from jackknife test were shown in Table 2.

The predictive accuracy of Naïve Bayes is approximately 3%, 4%, 5%, and 7% higher than that of the BayesNet, Random Forest, LogitBoot, and J48 classifiers, respectively. Although the accuracies of RBFnetwork and SVM are equal to that of Naïve Bayes, their auROC scores are lower than that of Naïve Bayes. These results indicate that the proposed Naïve Bayes model can be effectively used to classify phage virion and nonvirion proteins.

4. Conclusions

In this study, the Naïve Bayes classifier with feature selection method is presented to identity phage virion proteins based on the primary sequence information. By using Correlation-based Feature Subset Selection algorithm, the feature dimensions were reduced, and 38 prominent features that could remarkably improve the predictive accuracies were obtained. However, the detailed analyses of the selected features are required to provide more information about their roles in biological activity. The accuracy for the classification of phage virion and nonvirion proteins reached 79.15% in the jackknife test, indicating that the proposed method is an effective tool for phage virion protein identification. It is expected that the presented model will provide novel insights into the research on phage proteomics. Since user-friendly and publicly accessible web servers represent the future direction for developing practically more useful predictors [58], we shall make efforts in our future work to provide a web server for the method presented in this paper.

Acknowledgments

The authors wish to express their gratitude to three anonymous reviewers whose constructive comments were very helpful in strengthening the presentation of this paper. This work was supported by the National Nature Scientific Foundation of China (nos. 61100092, 61202256), the Fundamental Research Funds for the Central Universities (ZYGX2012J113).

References

[1] R. Lavigne, P. J. Ceyssens, and J. Robben, "Phage proteomics: applications of mass spectrometry," *Methods in Molecular Biology*, vol. 502, pp. 239–251, 2009.

[2] M. A. Hall, *Correlation-Based Feature Selection for Machine Learning: Data Mining, Inference and Prediction*, Springer, Berlin, Germany, 2008.

[3] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[4] K. C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, 2013.

[5] W. Z. Lin, J. A. Fang, X. Xiao, and K. C. Chou, "iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins," *Molecular BioSystems*, vol. 9, no. 9, pp. 634–644, 2013.

[6] X. Xiao X, P. Wang, W. Z. Lin, J. H. Jia, and K. C. Chou, "iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168–177, 2013.

[7] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, article e68, 2013.

[8] K. C. Chou, Z. C. Wu, and X. Xiao, "iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Molecular Biosystems*, vol. 8, no. 2, pp. 629–641, 2012.

[9] W. Chen, H. Lin, P. M. Feng, C. Ding, Y.-C. Zuo, and K.-C. Chou, "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS ONE*, vol. 7, no. 10, Article ID e47843, 2012.

[10] Y. Xu, J. Ding, L. Y. Wu, and K.-C. Chou, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS ONE*, vol. 8, no. 2, Article ID e55844, 2013.

[11] UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40, pp. D71–D75, 2012.

[12] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[13] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 43, no. 3, pp. 246–255, 2001.

[14] K. C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[15] L. Nanni, A. Lumini, D. Gupta, and A. Garg, "Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 2, pp. 467–475, 2012.

[16] D. Zou, Z. He, J. He, and Y. Xia, "Supersecondary structure prediction using Chou's pseudo amino acid composition," *Journal of Computational Chemistry*, vol. 32, no. 2, pp. 271–278, 2011.

[17] S. W. Zhang, Y. L. Zhang, H. F. Yang, C. H. Zhao, and Q. Pan, "Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies," *Amino Acids*, vol. 34, no. 4, pp. 565–572, 2008.

[18] K. K. Kandaswamy, G. Pugalenthi, S. Möller et al., "Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition," *Protein and Peptide Letters*, vol. 17, no. 12, pp. 1473–1479, 2010.

[19] S. Mei, "Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning," *Journal of Theoretical Biology*, vol. 310, pp. 80–87, 2012.

[20] Y. K. Chen and K. B. Li, "Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 318, pp. 1–12, 2013.

[21] M. Hayat and A. Khan, "Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC," *Protein and Peptide Letters*, vol. 19, no. 4, pp. 411–421, 2012.

[22] M. Khosravian, F. K. Faramarzi, M. M. Beigi, M. Behbahani, and H. Mohabatkar, "Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods," *Protein and Peptide Letters*, vol. 20, no. 2, pp. 180–186, 2012.

[23] H. Mohabatkar, M. M. Beigi, K. Abdolahi, and S. Mohsenzadeh, "Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach," *Medicinal Chemistry*, vol. 9, no. 1, pp. 133–137, 2013.

[24] M. M. Beigi, M. Behjati, and H. Mohabatkar, "Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach," *Journal of Structural and Functional Genomics*, vol. 12, no. 4, pp. 191–197, 2011.

[25] S. S. Sahu and G. Panda, "A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction," *Computational Biology and Chemistry*, vol. 34, no. 5-6, pp. 320–327, 2010.

[26] R. Zia-Ur and A. Khan, "Identifying GPCRs and their types with Chou's pseudo amino acid composition: an approach from multi-scale energy representation and position specific scoring matrix," *Protein and Peptide Letters*, vol. 19, no. 8, pp. 890–903, 2012.

[27] X. Y. Sun, S. P. Shi, J. D. Qiu, S. B. Suo, S. Y. Huang, and R. P. Liang, "Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform," *Molecular BioSystems*, vol. 8, no. 12, pp. 3178–3184, 2012.

[28] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653–660, 2008.

[29] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[30] H. Mohabatkar, "Prediction of cyclin proteins using Chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 17, no. 10, pp. 1207–1214, 2010.

[31] H. Mohabatkar, M. Mohammad Beigi, and A. Esmaeili, "Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[32] D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, and A. Torres, "Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 257, no. 1, pp. 17–26, 2009.

[33] B. Q. Li, T. Huang, L. Liu, Y. D. Cai, and K. C. Chou, "Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network," *PLoS ONE*, vol. 7, no. 4, Article ID e33393, 2012.

[34] T. Huang, J. Wang, Y. D. Cai, H. Yu, and K. C. Chou, "Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma," *PLoS ONE*, vol. 7, no. 4, Article ID e34460, 2012.

[35] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Analytical Biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[36] D. S. Cao, Q. S. Xu, and Y. Z. Liang, "Propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013.

[37] L. Montanucci, P. Fariselli, P. L. Martelli, and R. Casadio, "Predicting protein thermostability changes from sequence upon multiple mutations," *Bioinformatics*, vol. 24, no. 13, pp. i190–i195, 2008.

[38] M. M. Gromiha and M. X. Suresh, "Discrimination of mesophilic and thermophilic proteins using machine learning algorithms," *Proteins*, vol. 70, no. 4, pp. 1274–1279, 2008.

[39] L. C. Wu, J. X. Lee, H. D. Huang, B. J. Liu, and J. T. Horng, "An expert system to predict protein thermostability using decision tree," *Expert Systems with Applications*, vol. 36, no. 5, pp. 9007–9014, 2009.

[40] H. Lin and W. Chen, "Prediction of thermophilic proteins using feature selection technique," *Journal of Microbiological Methods*, vol. 84, no. 1, pp. 67–70, 2011.

[41] W. Chen and H. Lin, "Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine," *Computers in Biology and Medicine*, vol. 42, no. 4, pp. 504–507, 2012.

[42] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, NY, USA, 1997.

[43] J. Cao, R. Panetta, S. Yue, A. Steyaert, M. Young-Bellido, and S. Ahmad, "A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins," *Bioinformatics*, vol. 19, no. 2, pp. 234–240, 2003.

[44] M. Yousef, S. Jung, A. V. Kossenkov, L. C. Showe, and M. K. Showe, "Naïve Bayes for microRNA target predictions—machine learning for microRNA targets," *Bioinformatics*, vol. 23, no. 22, pp. 2987–2992, 2007.

[45] Y. Murakami and K. Mizuguchi, "Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites," *Bioinformatics*, vol. 26, no. 15, pp. 1841–1848, 2010.

[46] F. Sambo, E. Trifoglio, B. Di Camillo, G. M. Toffolo, and C. Cobelli, "Bag of Naïve Bayes: biomarker selection and classification from genome-wide SNP data," *BMC Bioinformatics*, vol. 13, supplement 14, article S2, 2012.

[47] K. C. Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space," *Proteins*, vol. 21, no. 4, pp. 319–344, 1995.

[48] G. P. Zhou, "An intriguing controversy over protein structural class prediction," *Journal of Protein Chemistry*, vol. 17, no. 8, pp. 729–738, 1998.

[49] K. C. Chou and D. W. Elrod, "Using discriminant function for prediction of subcellular location of prokaryotic proteins," *Biochemical and Biophysical Research Communications*, vol. 252, no. 1, pp. 63–68, 1998.

[50] G. P. Zhou and N. Assa-Munt, "Some insights into protein structural class prediction," *Proteins*, vol. 44, no. 1, pp. 57–59, 2001.

[51] Y. X. Pan, Z. Z. Zhang, Z. M. Guo, G. Y. Feng, Z. D. Huang, and L. He, "Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach," *Journal of Protein Chemistry*, vol. 22, no. 4, pp. 395–402, 2003.

[52] G. P. Zhou and K. Doctor, "Subcellular location prediction of apoptosis proteins," *Proteins*, vol. 50, no. 1, pp. 44–48, 2003.

[53] K. C. Chou and H. B. Shen, "Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms," *Natural Science*, vol. 2, no. 10, pp. 1090–1103, 2010.

[54] M. Hayat and A. Khan, "MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM," *Journal of Theoretical Biology*, vol. 292, pp. 93–102, 2012.

[55] X. Xiao, Z. C. Wu, and K. C. Chou, "iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites," *Journal of Theoretical Biology*, vol. 284, no. 1, pp. 42–51, 2011.

[56] K. C. Chou, Z. C. Wu, and X. Xiao, "iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins," *PLoS ONE*, vol. 6, no. 3, Article ID e18258, 2011.

[57] H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2005.

[58] K. C. Chou and H. B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes," *Natural Science*, vol. 2, pp. 63–92, 2009.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 653759, 12 pages
http://dx.doi.org/10.1155/2013/653759

Review Article

Uses of Phage Display in Agriculture: A Review of Food-Related Protein-Protein Interactions Discovered by Biopanning over Diverse Baits

Rekha Kushwaha,[1,2] Christina M. Payne,[3,4] and A. Bruce Downie[2,5]

[1] *Department of Horticulture, Agricultural Science Center North, University of Kentucky, Room 308J, Lexington, KY 40546, USA*
[2] *Seed Biology Group, University of Kentucky, Lexington, KY 40546, USA*
[3] *Department of Chemical and Materials Engineering, University of Kentucky, Room 159, F. Paul Anderson Tower, Lexington, KY 40546, USA*
[4] *Center for Computational Sciences, University of Kentucky, Lexington, KY 40506, USA*
[5] *Department of Horticulture, University of Kentucky, Room 401A, Plant Science Building, Lexington, KY 40546, USA*

Correspondence should be addressed to A. Bruce Downie; adownie@uky.edu

Received 27 February 2013; Accepted 2 April 2013

Academic Editor: Jian Huang

This review highlights discoveries made using phage display that impact the use of agricultural products. The contribution phage display made to our fundamental understanding of how various protective molecules serve to safeguard plants and seeds from herbivores and microbes is discussed. The utility of phage display for directed evolution of enzymes with enhanced capacities to degrade the complex polymers of the cell wall into molecules useful for biofuel production is surveyed. Food allergies are often directed against components of seeds; this review emphasizes how phage display has been employed to determine the seed component(s) contributing most to the allergenic reaction and how it has played a central role in novel approaches to mitigate patient response. Finally, an overview of the use of phage display in identifying the mature seed proteome protection and repair mechanisms is provided. The identification of specific classes of proteins preferentially bound by such protection and repair proteins leads to hypotheses concerning the importance of safeguarding the translational apparatus from damage during seed quiescence and environmental perturbations during germination. These examples, it is hoped, will spur the use of phage display in future plant science examining protein-ligand interactions.

1. Introduction

Since its development by Smith [1], phage display has proven to be a powerful tool for protein interaction studies in Immunology, cell biology, drug discovery, and pharmacology. Phage display is one of the preeminent means by which scientists identify proteins having affinity for other molecules and has a staggering throughput capacity for screening with libraries with titers approaching 10^9 virions per microliter. Its utility lies principally in generating molecular probes against specific targets and for the identification, analysis, and manipulation of protein-ligand (including protein-protein) interactions. Modern phage display libraries permit the sought attribute (namely, protein with affinity for a ligand (bait)) to be directly coupled to the DNA sequence encoding the protein in a nondestructive manner. Random DNA libraries, or those formed from cDNA after randomly priming mRNA, provide a host of different amino acid contexts that can translate into a continuum of affinities for the bait. Recovery of overlapping clones of a particular protein permits examination of this region of the protein, directing the experimenter to the specific site capable of binding the ligand. With the protein-binding site effectively located, this information can be used to predict target attributes that serve

as the foundation of ligand-protein affinity, guiding future protein engineering efforts.

This technique, due to its simplicity and efficacy, has been responsible for discoveries of synthetic antibodies and molecular interactions and utilized in directed evolution. The applications of phage display for discovery of protein-ligand interactions have become increasingly complex as its utility has been recognized in a diversity of fields, including the identification of targets of bioactive molecules. For example, Huperzine A is a plant-produced, bioactive compound with multiple neuroprotective effects [2, 3]. Magnetic biopanning approaches have been used to identify some of the target pathways influenced by Huperzine A's pharmacological effects which are responsible for alleviating a host of dysfunctions, potentially including Alzheimer's disease [4].

Despite the utility of phage display, the technique has received less attention from plant scientists, with the exception of sustained programs developing antibodies to a host of different cell wall components [5], a topic discussed in other literature [6] and thus not examined here. However, phage display has much to offer other fields of plant research. This review surveys the applications of phage display in the discovery of protein-protein interactions in various fields of plant science concerned with maximizing crop plants' seed production and the utilization of the nutrients stored in seeds, from protecting crops from harmful pests to alleviating human allergenic reactions to seed storage proteins.

Our objective in highlighting this literature is to heighten the awareness of plant biologists to the utility of the technique for more than antibody production alone. If successful, phage display should figure more prominently in the research of those plant scientists examining molecular interactions in the future.

2. Applications of Phage Display in Agriculture: Seed Production

Why focus on seed production? On a fundamental level, it is necessary to understand seed attributes as human reliance on seeds is so pervasive. Seeds are our major food source (70% of our diet [7, 8]); they are fodder for our livestock, a method of bulk food transport, storage, germplasm preservation, and a vehicle for technology delivery. It is imprudent not to understand more about how a seed fulfills its function as a propagule, a process on which we depend so utterly, yet about which we still know so very little [9, 10]. In addition to constituting the majority of humanity's food, recent additional uses for the energy stored in seeds (biofuels [11]) have periodically led to higher seed and commodity prices worldwide [12, 13]. While governments attempt to mitigate the negative impact of increasing staple food prices on the poor [12], demand for seed as food and biofuel feedstock and the land on which to produce it continues to increase [14]. The growing global population is projected to increase cereal consumption for food alone by a billion metric tons in the next 30 years (FAO, 2002, http://www.fao.org/docrep/004/y3557e/y3557e00.htm); yet yield losses due to unpredictable biotic and abiotic stresses are projected to increase [15]. These grim facts have added urgency to the requirement to improve understanding of all facets of seed production. It is imperative that we do this if we are to feed ourselves [16].

2.1. Phage Display Utilized in the Defense of Plants against Herbivores and Microbes

2.1.1. Identification and Production of Superior Protease Inhibitors. Protease inhibitors (PIs) are one defense system plants employ against herbivores and microorganisms [17]. PIs are a plant protection strategy that can attenuate nutrient assimilation in the insect gut or by microbes by inhibiting the activity of pest digestive proteases [18]. There are a large number of PIs used by plants as natural protection against pests [19]. PI production can be induced in the plant body by pest/pathogen attack through the jasmonic acid pathway [20], but are also subject to developmental regulation, their production being stimulated in storage tissues [21]. In seeds, PI transcription is stimulated by abscisic acid (ABA) (inhibitory to germination) and inhibited by gibberellic acid (GA) (stimulatory for germination) [22]. Thus, endogenous seed protease activity (responsible for storage protein breakdown for use by the establishing seedling) is reduced during the anabolic period of seed development, permitting unhampered accumulation of the storage proteins, while this hindrance is alleviated during the period of seedling establishment allowing access to energy and components constituting the storage proteins (Figure 1(a)). Reduction, through the NADPH-dependent thioredoxin h system, of specific disulfide bonds necessary to impart the PI with its inhibitory confirmation [23] also aids the removal of seed PI influence from establishing seedlings [24]. Typically, PIs are heat labile, permitting humans to acquire the full nutritional value of the seed storage proteins (some of which are protease inhibitors in their own right [25]) in cooked food that is denied to insects and microorganisms [26].

The plant usually encodes a considerable variety of PIs that are used to inhibit a wide range of pest proteases and isoforms within a protease class. Protease isoform prevalence in the insect can vary, exhibiting adaptability on the part of the pest in attempts to overcome this plant defensive mechanism [31–34]. Strategies using phage display to inform directed evolution [35, 36] or specific site-directed mutation [37] efforts to produce PIs with greater specificity [38] or affinity [39] for the pest protease active site aim at enhancing this natural means of protecting crops. The PIs are usually quite specific for their protease target [40], and phage display has been at the center of efforts to construct PIs with a greater range of targets. This enhanced generality includes biopanning for PI variants that can inhibit proteases of a diversity of insect pests [41]. Another facet of phage display-based protection enhancement takes the opposing strategy, endeavoring to identify PIs that are even more finely tuned to the target species (pest) protease class [42].

These various attempts to use phage display to acquire novel PIs are geared toward providing a greater range of PIs

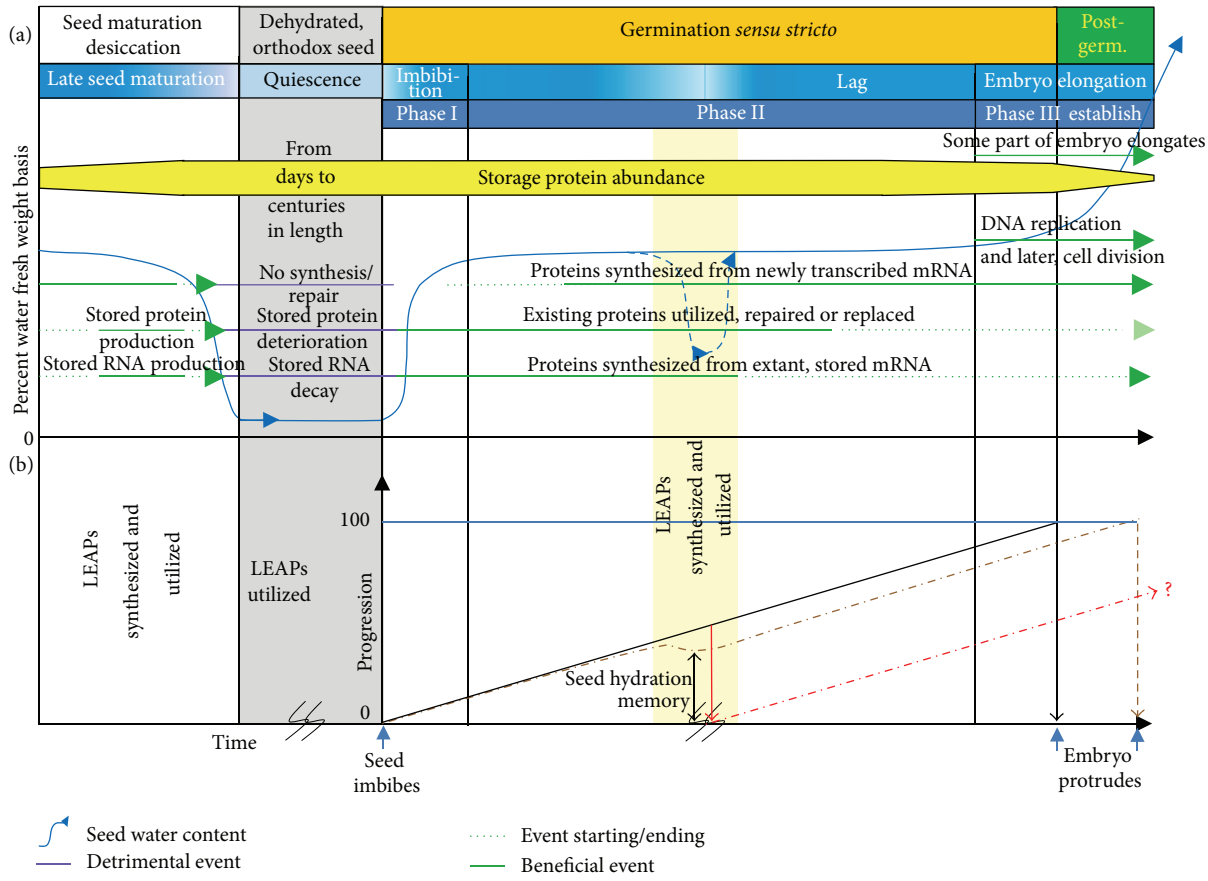


Figure 1: A graphic depiction of events occurring during the stages of late maturation, quiescence, and germination of orthodox seeds [27]. (a) Four stages during a plant's lifecycle commencing with seed maturation desiccation and ending with postgermination seedling establishment (Postgerm). Seed water content is represented by the solid blue line in the graph and is depicted as well by shades of blue in the background highlighting stages in the continuum encompassing late seed maturation, quiescence, and the three classical phases of water uptake during seed germination (imbibition, lag, and embryo elongation/seedling establishment (establish)). Phase III has been placed to span the completion of germination because turgor-driven embryo cell expansion, required to protrude from the seed, necessitates additional water uptake. The axis representing time has been broken during quiescence to emphasize that, although this period can last for centuries, certain species seeds remain viable [28, 29]. Events that are beneficial for the preparation of maturation desiccation or the resumption of growth are presented as green lines. Events occurring that are detrimental to the cellular constituents are depicted as purple lines. The commencement and termination of these events are signified by short-dashed lines. A drying event, followed by rehydration during germination, has been inserted as a long-dashed blue line. This region is also highlighted by yellow shading that depicts a period of high temperature stress. The abundance of the seed storage proteins is depicted as a yellow bar whose thickness is tapered at both ends to signify net accumulation during late embryogenesis and rapid hydrolysis during seedling establishment. (b) Late embryogenesis abundant protein (LEAP) synthesis and utilization during late seed maturation and quiescence. The overall progression of a non-dormant (quiescent) seed toward the completion of germination (100% progression) is depicted as a solid line commencing at the arrow (seed imbibes) on the time axis. To emphasize the capacity of the seed to preserve its physiology at a point above 0 progression (y-axis) during the dehydration/supraoptimal temperature event (dash-dotted brown line), the trajectory of progression deviates partially from that had no drying/thermal stress occurred. The red line, and the dash-dotted red progression trajectory emanating from it, portrays a seed without the capacity to preserve its physiology. The difference (double-headed arrow) is the seed hydration memory [30]. The only manifestation of the stressful event interrupting the progression of germination is a slightly delayed point on the time axis at which the embryo protrudes. A seed unable to maintain its physiology may or may not be capable of completing germination, hence the question mark. The production of the LEAPs and their utilization to presumably preserve the seed's physiology, post-imbibition, are indicated. The time axis is broken during the stressful event to signify its unknown duration. Graph adapted from Nonogaki et al. [10].

affording protection to plants than is available to the conventional plant breeder. The development and identification of PIs with unique capabilities of downregulating the activity of specific pest proteases, through phage display or other means, will permit these plant protection mechanisms to augment those existing naturally in the plant. Stacking PIs with different protease target sites may help to broaden pest susceptibility while delaying the acquisition of resistance to the PIs [43, 44].

2.1.2. Discovering Non-Protease Inhibitor Protective Peptides. Phage display can identify peptides or proteins that have

affinity for a vast array of molecules. Peptides with high affinity for proteins key to a pest's lifecycle can be disruptive to the pest's permanence or pathology [45]. Once identified, such peptides can be engineered and introduced into most crop plants for endogenous production, providing a novel line of defense against plant pests. Such specific, plant-contained, protective mechanisms may prove to be less damaging to off-target organisms in the crop environment than conventional pesticides [46]. Chemoreception-disruptive peptides selected from peptide libraries have been shown to decrease parasitism by nematodes, albeit at doses 3 orders of magnitude greater than the Aldicarb nematocide control [47]. Despite this much lower competence, the Aldicarb mimetic with high affinity to acetylcholinesterase, when produced *in planta*, was effective in reducing parasite load in potato by cyst nematodes [46] that are otherwise difficult to control due to their sessile habit and location, embedded in the plant roots. Therefore, *in situ* production of the mimetic with a lower efficacy counteracted this liability, resulting in nematode control, which is also the goal of the generally applied nematocide possessing greater potency but only a portion of which arrives at the site of action. Similarly, phage display identified peptides binding to zoospores of the fungal pathogen *Phytophthora capsici*. Many of the zoospore-binding peptides resulted in the premature encystment of the zoospore without any other inductive signal. In addition to aiding in the identification of zoospore-displayed receptors controlling encystment, the authors postulated that such peptides might represent a novel plant defensive mechanism [48]. Subsequently, decreased infection by this soil-borne fungus resulted when a protective peptide was expressed *in planta* in a form allowing its secretion into the rhizosphere [49].

2.1.3. Uses in Plant Virology. Phage display has been used by various plant virologists in identification of peptides that bind to a pathogenic virus's coat protein. The phage display-isolated peptides were very specific and highly sensitive. At the very least, these have diagnostic potential as they can be produced as fusions with proteins that serve as an antigen for antibody-reporter molecule conjugates [50]. They may also constitute the basis for a novel, introduced disease resistance strategy. Peptides with high affinity and specificity for vital viral proteins could be identified, and subsequently, the capacity to synthesize these peptides may be introduced into plants. *In planta* peptide production might prevent viral proliferation in infected cells. Such a strategy has been used successfully with antibodies [51], but antibody folding usually requires an oxidizing environment conducive to forming specific intracellular disulfide bonds necessary for function [52]. Phage display-selected peptides may not be so exacting in their requirements [53]. Indeed, phage display-selected peptides capable of binding to a coat protein of the rice black streaked dwarf virus (RBSDV), when produced recombinantly for diagnostic purposes, have been shown to also disrupt proper coat protein folding and reduce the pathogenicity of RBSDV [54]. Phage display has also assisted in the elucidation of various host systems secunded to the virus to permit successful infection and replication. Using the viral replication enhancer protein, AC3 as bait, a phage library of random dodecapeptides fused to a coat protein was panned to identify interacting peptides that were then analyzed for homology to proteins from the model plant, *Arabidopsis thaliana*. The revelation of the pathways to which these proteins are integral has allowed a more sophisticated understanding of events required for successful viral lifecycle and the role of the multifunctional protein AC3 in events leading to virus-induced gene silencing [55].

2.1.4. Identification of Immune Targets in Plants. Plants are known to have a very complex and diverse immune system against microbes [56]. The first active line of defense occurs at the plant cell surface when microorganism-associated molecular patterns (MAMPs) such as lipopolysaccharides, peptidoglycans, or bacterial flagellin are detected by pattern recognition receptors (PRRs). These PRRs are responsible for pattern-triggered immunity (PTI) in plants [57–59]. To circumvent PTI, adapted pathogens can deliver effector molecules directly into the plant cell. As a countermeasure, plants have developed corresponding resistance (R) proteins to recognize these effectors and their modified targets which results in effector-triggered immunity (ETI) [59]. Both PTI and ETI involve specific families of proteins but the distinction between both types is not yet clear. What is clear is that a large number of proteins participate in the immunity process. Rioja et al. used phage display to study these interactions and to identify *Arabidopsis* proteins able to bind bacterial pathogens [60]. For this, they constructed two phage-display libraries from the cDNA of microbe-challenged *Arabidopsis*. Recombinant phage displaying plant proteins capable of interacting with different species of *Pseudomonas* (the pathogen) were selected by biopanning using microbial cells as selection ligands. In this way, plant proteins involved in defense responses were identified and subsequently confirmed *in vitro* for the capacity to bind microbial cells. Using different strains of *Pseudomonas* as bait allowed discrimination between common bacterial receptors and specific targets of virulent or avirulent strains.

2.2. Applications in Cell Wall Research. Interest in using cellulose and other plant cell wall components as feedstock for biofuel production continues to grow worldwide for a host of reasons. Current means of deconstructing cellulose polysaccharides to glucose for conversion to biofuels are less efficient and more expensive than practical for an industrially relevant process. One avenue being explored for more efficient conversion of cellulose to glucose is through enhanced enzymatic degradation. It has been demonstrated that some cellulases and hemicellulases retain their function when fused to a viral coat protein [61, 62]. These clones can subsequently be reengineered to alter (randomize) specific regions of interest imparting novel functionalities/affinities to the displayed enzyme combinatorially. The resultant library of phage displayed variant enzymes can then be screened over substrates/inhibitors to study the individual amino acids imparting the observed/desired property.

Programs have also used phage display libraries to discover or improve upon carbohydrate binding modules focused on the use of these regions to enhance the binding affinity of the glycoside hydrolase/binding module construct to various crystalline morphologies, which may improve upon their productivity [63]. Additional uses include highly specific probes for cell wall constituents, which are critical to refining our understanding of plant cell wall construction [64–66].

Furthermore, a library of fungal endo-β-1,4-xylanase enzyme variants permitted the simultaneous assessment of the influence of many different individual residues on the affinity for xylanase inhibitor proteins [67]. Subsequent work has permitted the development of an endo-β-1,4-xylanase enzyme that retains its catalytic competence while being completely insensitive to xylanase inhibitor proteins found in wheat flour [68]. The fungal xylanase is used in the food industry to enhance nutritional value and properties, but its inactivation by the endogenous inhibitors found in the foodstuffs on which it is used has been a problem for the industry. Moreover, through a computational approach, the pH stability of the enzyme has now been greatly improved leading to an increase in its utility in the food preparation industry [69].

2.3. Phage Display Uses in Combating Allergies to Seed Storage Proteins. Almost 5% of humans have some form of food hypersensitivity [70]. Identified food allergens include the seed storage proteins that can induce a variety of allergic syndromes [71, 72]. Phage display has assisted in the rapid identification of antigens eliciting hypersensitive responses [73] including those previously uncataloged [74]. Once individuals suspect they are allergic to a particular food, a more sophisticated assessment of the component(s) in the food causing the allergic reaction is necessary if any alleviation is to be attained. Epitopes from a library of allergens from the food in question [75], panned over patient IgE, can rapidly and cheaply identify the specific allergen(s) causing the hypersensitive response [74]. For example, peanut allergies are quite common (~1% of the population of the USA [76]), are perceived to be increasing [77], and can be severe [78]. Phage display has been used to identify precisely what proteins are causing the hypersensitive reaction in peanut-sensitive patients [79], implicating the seed storage proteins as significant and accounting for 6 of the 8 allergens identified in peanut to date [80].

Similarly, "baker's asthma," a common occupational affliction, was until recently only known to be caused by an allergic reaction to "flour" components. Phage display was used to identify a causal agent in wheat flour as native gliadin (33% of all cases) and, more specifically, α- and β-gliadin, which were causal in 12% of all Baker's asthma [81]. The use of such epitope display accurately identifies the causal agent of food allergies that, once identified, can be the subject of investigations aimed at rendering it less antigenic. Such an approach has been used in a program aimed at mitigating allergenic reactions in celiac disease.

Celiac (or also coeliac) disease affects approximately 1% of the human population [82]. It is induced by components in several cereal storage proteins in common use (bread, pasta, and beer). It is a complex disease with aspects of both autoimmune disease and food hypersensitivity [83]. In the autoimmune response, tissue transglutaminase (tTG) enzyme is targeted by self-antibodies but only after gluten ingestion when tTG is complexed with gluten [84, 85]. The enzyme deaminates the abundant glutamine residues, which can comprise up to ~35–40% of the amino acids constituting the α-gliadin component of gluten [86]. Antibodies are also specifically produced against tTG-deaminated gliadin fragments from gluten, a hallmark of food hypersensitivity [87].

Approaches to alleviate disease symptoms include attempts to block portions of gliadin using synthetic, high-affinity peptides, thus preventing tTG action/gliadin modification and subsequent formation of immunostimulatory epitopes. Phage display has played a critical role in the identification of the peptides possessing a strong affinity for gliadin. These act to first depress tTG activity against the gliadin substrate *in vitro* by steric hindrance, the eventual goal being to attenuate the autoimmune response by decreasing the association of the enzyme with its substrate, minimizing inflammation *in vivo* [88]. The second prong of this program is to cover the epitopes on gliadin, masking the protein fragments from the antibodies binding to them [89]. This program has passed the first several hurdles in the long road to providing a modicum of relief for celiac disease sufferers, including proof that the synthetic peptides act to block tTG activity against gliadin as did the phage-tethered peptides on which they were based, which does not necessarily follow [90]. The program awaits trials of the identified gliadin-binding peptides *in vivo*. In addition to their potential therapeutic uses, the various peptides, binding to different sites on the gliadin protein, [89] could provide valuable tools for researchers in the field of celiac disease.

2.4. Phage Display Identifies Protein Isoaspartyl Methyltransferase Substrates in the Stored Seed Proteome. The tTG-mediated alteration of gliadin glutamine residues, through deamidation, enhanced the antigenicity of gliadin fragments [91]. The proteins present in dry seeds are particularly susceptible to a host of nonenzymatic conversions, many of which are deleterious [92–99], and some of which may play a role in preparing the seed for the completion of germination upon rehydration [100]. Regardless, these conversions can also result in peptides that are recognized by the human immune system or are recalcitrant to hydrolysis. For example, spontaneous isoaspartyl formation is known to result in autoimmune responses [101] and interfere with peptide degradation [102] decreasing the nutritional value of ingested seed products [103] and, if sufficiently widespread in the stored proteome, would be disastrous for germination and seedling establishment [104, 105].

Orthodox seeds [27] are capable of extreme dehydration allowing them to remain viable in extremes of temperature [106, 107] and in some instances, for centuries [28, 29]. This

remarkable feat means that the seed proteome is at risk for deleterious alteration for the whole of this time as there is insufficient water present to effect repair. A prominent detrimental alteration is the conversion of L-Asn or L-Asp residues in proteins to succinimide that, upon water addition, usually converts to the unusual, uncoded amino acid, L-isoAsp [108–111]. In the imbibed state, isoAsp in proteins is recognized, methylated, and repaired by protein L-isoaspartyl methyltransferase (PIMT) [112, 113].

What proteins are most at risk for isoAsp formation or for which PIMT has highest affinity? Due to the labile nature of the labeled isoAsp and susceptibility of proteins to form isoAsp during rigorous extraction necessary to obtain samples, these identifications have not been facile [114–116]. Moreover, the abundance and susceptibility to damage of the seed storage proteins [93] have made identification of additional PIMT target proteins using extracts from seeds difficult [116].

An alternative approach used phage display to mitigate the influence of protein extraction on the generation of isoAsp while largely removing the seed storage proteins from the analysis [117, 118]. A group of proteins involved in aspects of translation were revealed as important substrates of PIMT in seeds. This led to the realization that the stored proteins essential for the translational apparatus must be especially important to protect from general dysfunction because there is no means of replacing them (or any other protein) from either the stored or *de novo* produced transcriptomes if translation is compromised in the majority of cells comprising a tissue and/or organelles [119, 120] present in cells (Figure 2).

2.5. Phage Display Identifies Late Embryogenesis Abundant Protein Client Proteins in the Seed. One of the targets recovered from the biopans over PIMT1 and not directly involved in translation was the seed maturation protein1 (SMP1; At3G12960), a Pfam (PF04927) SMP late embryogenesis abundant (LEA) protein homolog to the soybean (*Glycine max*) SMP, *Gm*PM28 (Glyma08G18400). LEA proteins [121, 122] are thought to assist anhydrobiosis (life without water), an attribute of many microorganisms, lichens, and some animals and plants [123–136]. This trait has underpinned agriculture for millennia [137, 138], allowing a portion of each seed harvest to be withheld, dehydrated, and hence, resistant to pathogen attack, and to establish the next crop, either the subsequent year or decades into the future [139].

The recovery of an LEA protein by PIMT1 was intriguing as it may indicate that this LEA protein needs protection from isoAsp formation by PIMT1 to retain its function, forming part of an interactive network of protein protective mechanisms extant in seeds. T-DNA insertional mutants of this LEA in two different *Arabidopsis* ecotypes were incapable of entering secondary dormancy when seeds were exposed to supraoptimal (40°C) germination temperatures for several days prior to being placed at permissive temperatures (25°C) [117]. Such a specific phenotypic manifestation of the loss of this LEA's function suggested it safeguarded a crucial subset of proteins involved in the proteomic memory of environmental conditions the seed has experienced thus far following imbibition (supraoptimal temperatures). High temperature and/or desiccation after a period of imbibition during which important environmental cues had been perceived and the transcriptome/proteome altered accordingly, but prior to radicle protrusion, would expose the proteome and the integrated environmental information it represents to deleterious conditions. This necessitates protective mechanisms be invoked to ensure the heat-stressed/dehydrated proteins retain their function so that germination can resume at the appropriate point at which it left off once the seeds are rehydrated [140]. Dubrovsky [30] referred to the capacity of seeds to resume germination from the point at which they had progressed prior to dehydration as the "seed hydration memory" (Figure 1(b)).

The concept of the LEA proteins safeguarding environmental cues, acquired during the imbibed period and embodied in a heat-sensitive proteome, can be subsumed into their role of aiding the survival of water loss during maturation desiccation, quiescence or after imbibition [141, 142]. The dysfunction of some heat-labile molecule(s), when not protected by SMP1, results in a seed that cannot "remember" the supraoptimal temperature it has experienced and thus behaves inappropriately, completing germination immediately when removed to 25°C rather than entering thermal dormancy (Figure 1(b)).

It was necessary to ascertain with what target proteins the SMP1 LEA protein associates because these would be candidates for controlling the induction of secondary dormancy due to high heat [117] but this was not known. In fact, uncertainty exists regarding whether LEA proteins serve exclusively as general "spacer" molecules ("molecular shields" or crowders) that simply prevent deleterious aggregation upon water loss or if they can act as specific protectors of individual target molecules so-called "client molecules" [143–145]. Therefore, recombinant SMP1 and its soybean *Gm*PM28 homolog were used as bait in screens at two different temperatures and with two independently produced Arabidopsis seed, phage display libraries [146]. Biopanning over these recombinant LEA homologs demonstrated that the same protein clients, indeed the same region of the same protein clients, are consistently retrieved by both baits at two different temperatures [146]. The client proteins identified did not have a single target protein in common with the PIMT1 screens, yet those involved in translation were again prominent among the protected target proteins further entrenching the contention that protection of the proteins involved in translation is paramount for safeguarding the longevity of orthodox seeds (Figure 2).

3. Conclusions

Predictions of dire consequences for humanity if food (read seed) production is not drastically increased is a goad for researchers investigating seed production to endeavor to understand more of the complexities of this event. Frequently, the understanding sought lies at the level of protein-ligand or protein-protein interactions. In this regard, phage display has proved extremely useful for both the discovery of such

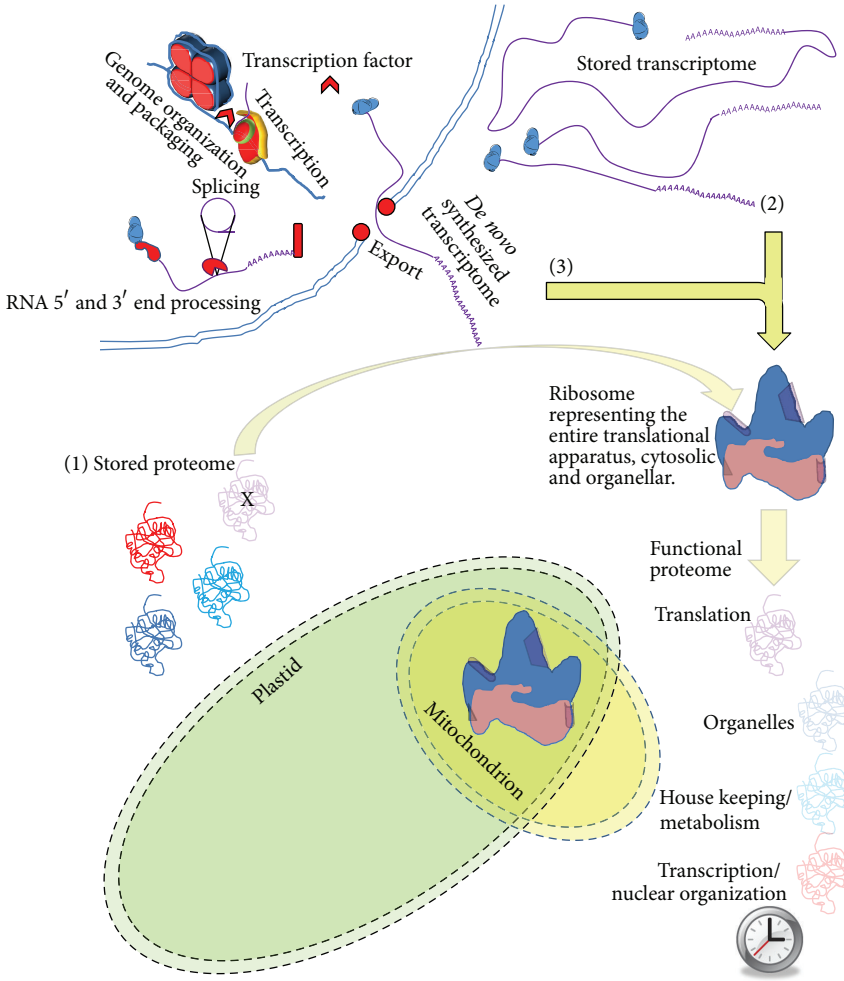


Figure 2: Those proteins essential to translation are the proteome's "Achilles' heel" for seed longevity. In the imbibed seed, there are three means by which functional proteins can be recruited into the newly reestablished, active metabolism. The proteins may be part of (1) the stored proteome that has survived maturation desiccation and subsequent rehydration with their function intact. New protein can be translated from either (2) the stored transcriptome consisting of mRNA, produced during seed maturation, that survived maturation desiccation/rehydration or (3) *de novo* transcribed mRNA. Only those proteins essential to translation *must* be present in the stored proteome, sufficiently numerous and in an active state following imbibition, to carry out translation (probably with an emphasis on self-replacement) if the embryo is to survive. Various classes of proteins are color coded according to their function (red: transcription/nuclear organization; light blue: House-keeping/metabolism; dark blue: organelles; purple: translation). The proteins essential to translation are depicted decorating the ribosome in the cytosol, or in those organelles with their own genomes. The dysfunction of the proteins essential for translation has been emphasized by their partial transparency and an "X" through the molecule representing this class in the stored proteome. A lack of translation results in the eventual demise of the entire proteome over time (partially transparent functional proteome).

interactions and their subsequent manipulation towards an end. This review has highlighted, for the first time, the impact phage display has had on agricultural research concerned with seed production. Efforts to safeguard the crop plant's capacity to produce seeds and to protect the seeds themselves for exclusive human use/consumption have successfully employed phage display. Phage display has aided in the production of enzymes specialized for use in food processing, making nutrients more readily available. It has also provided the means of specifically identifying the causal agent(s) of seed allergies, and indications are that it may be instrumental in providing the first means of mitigating the effects of a prominent seed-related ailment. The use of phage display has permitted insights into the seed's endogenous natural protective and repair mechanisms, allowing a more fundamental understanding of the events transpiring during late embryogenesis, quiescence, and germination; in short, what makes seeds so excellent in their role as propagules.

Acknowledgments

This project was partially funded by an NSF IOS (0849230), Hatch, McIntire-Stennis (AD421 CRIS), USDA Seed Grant (2011-04375), and Sir Frederick McMaster Research Fellowship to ABD.

References

[1] G. P. Smith, "Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface," *Science*, vol. 228, no. 4705, pp. 1315–1317, 1985.

[2] R. Wang and C. T. Xi, "Neuroprotective effects of huperzine A: a natural cholinesterase inhibitor for the treatment of Alzheimer's disease," *NeuroSignals*, vol. 14, no. 1-2, pp. 71–82, 2005.

[3] Z. F. Wang, J. Wang, H. Y. Zhang, and X. C. Tang, "Huperzine A exhibits anti-inflammatory and neuroprotective effects in a rat model of transient focal cerebral ischemia," *Journal of Neurochemistry*, vol. 106, no. 4, pp. 1594–1603, 2008.

[4] W. Guo, S. Liu, J. Peng et al., "Examining the interactome of huperzine A by magnetic biopanning," *Public Library of Science ONE*, vol. 7, no. 5, article e37098, 2012.

[5] W. G. T. Willats, C. G. Steele-King, L. McCartney, C. Orfila, S. E. Marcus, and J. P. Knox, "Making and using antibody probes to study plant cell walls," *Plant Physiology and Biochemistry*, vol. 38, no. 1-2, pp. 27–36, 2000.

[6] A. J. Bernal and W. G. T. Willats, "Plant science in the age of phage," *Trends in Plant Science*, vol. 9, no. 10, pp. 465–468, 2004.

[7] J. D. Bewley and M. Black, *Seeds: Physiology of Development and Germination*, Plenum Press, New York, NY, USA, 1994.

[8] J. D. Bewley, K. J. Bradford, H. W. M. Hilhorst, and H. Nonogaki, *Seeds. Physiology of Development, Germination and Dormancy*, Springer, New York, NY, USA, 3rd edition, 2013.

[9] J. D. Bewley, "Seed germination and dormancy," *The Plant Cell*, vol. 9, no. 7, pp. 1055–1066, 1997.

[10] H. Nonogaki, G. W. Bassel, and J. D. Bewley, "Germination-still a mystery," *Plant Science*, vol. 179, no. 6, pp. 574–581, 2010.

[11] E. Kintisch, "Sowing the seeds for high-energy plants," *Science*, vol. 320, no. 5875, article 478, 2008.

[12] L. P. Koh and J. Ghazoul, "Biofuels, biodiversity, and people: understanding the conflicts and finding opportunities," *Biological Conservation*, vol. 141, no. 10, pp. 2450–2460, 2008.

[13] R. A. Sedjo, "Biofuels: think outside the cornfield," *Science*, vol. 320, no. 5882, pp. 1420–1421, 2008.

[14] J. Fargione, J. Hill, D. Tilman, S. Polasky, and P. Hawthorne, "Land clearing and the biofuel carbon debt," *Science*, vol. 319, no. 5867, pp. 1235–1238, 2008.

[15] J. L. Hatfield, K. J. Boote, B. A. Kimball et al., "Climate impacts on agriculture: implications for crop production," *Agronomy Journal*, vol. 103, no. 2, pp. 351–370, 2011.

[16] J. Fernandez-Cornejo, *The Seed Industry in U.S. Agriculture: An Exploration of Data and Information on Crop Seed Markets, Regulation, Industry Structure, and Research and Development*, U.S. Department of Agriculture, Economic Research Service, Washington, DC, USA, 2004.

[17] C. A. Ryan, "Protease inhibitors in plants—genes for improving defenses against insects and pathogens," *Annual Review of Phytopathology*, vol. 28, pp. 425–449, 1990.

[18] H. Koiwa, R. E. Shade, K. Zhu-Salzman et al., "Phage display selection can differentiate insecticidal activity of soybean cystatins," *The Plant Journal*, vol. 14, no. 3, pp. 371–379, 1998.

[19] F. de Leo, M. Volpicella, F. Licciulli, S. Liuni, R. Gallerani, and L. R. Ceci, "PLANT-Pls: a database for plant protease inhibitors and their genes," *Nucleic Acids Research*, vol. 30, no. 1, pp. 347–348, 2002.

[20] H. Koiwa, R. A. Bressan, and P. M. Hasegawa, "Regulation of protease inhibitors and plant defense," *Trends in Plant Science*, vol. 2, no. 10, pp. 379–384, 1997.

[21] L. Pouvreau, H. Gruppen, S. R. Piersma, L. A. van den Broek, G. A. van Koningsveld, and A. G. Voragen, "Relative abundance and inhibitory distribution of protease inhibitors in potato juice from cv. Elkana," *Journal of Agricultural and Food Chemistry*, vol. 49, no. 6, pp. 2864–2874, 2001.

[22] S. E. Jacobsen and N. E. Olszewski, "Gibberellins regulate the abundance of RNAs with sequence similarity to proteinase inhibitors, dioxygenases and dehydrogenases," *Planta*, vol. 198, no. 1, pp. 78–86, 1996.

[23] C. Marx, J. H. Wong, and B. B. Buchanan, "Thioredoxin and germinating barley: targets and protein redox changes," *Planta*, vol. 216, no. 3, pp. 454–460, 2003.

[24] F. Montrichard, F. Alkhalfioui, H. Yano, W. H. Vensel, W. J. Hurkman, and B. B. Buchanan, "Thioredoxin targets in plants: the first 30 years," *Journal of Proteomics*, vol. 72, no. 3, pp. 452–474, 2009.

[25] S. Candido Ede, M. F. Pinto, P. B. Pelegrini et al., "Plant storage proteins with antimicrobial activity: novel insights into plant defense mechanisms," *The Federation of American Societies for Experimental Biology Journal*, vol. 25, no. 10, pp. 3290–3305, 2011.

[26] S. N. Nahashon and A. K. Kilonzo-Nthenge, "Advances in Soybean and Soybean by-products in monogastric nutrition and health," in *Soybean and Nutrition*, P. H. El-Shemy, Ed., pp. 125–156, InTech, Shanghai, China, 2011.

[27] E. H. Roberts, "Predicting the storage life of seeds," *Seed Science and Technology*, vol. 1, pp. 499–514, 1973.

[28] J. Shen-Miller, M. B. Mudgett, J. W. Schopf, S. Clarke, and R. Berger, "Exceptional seed longevity and robust growth: ancient Sacred Lotus from China," *American Journal of Botany*, vol. 82, no. 11, pp. 1367–1380, 1995.

[29] S. Sallon, E. Solowey, Y. Cohen et al., "Germination, genetics, and growth of an ancient date seed," *Science*, vol. 320, no. 5882, p. 1464, 2008.

[30] J. G. Dubrovsky, "Seed hydration memory in Sonoran Desert cacti and its ecological implication," *American Journal of Botany*, vol. 83, no. 5, pp. 624–632, 1996.

[31] M. A. Jongsma, P. L. Bakker, J. Peters, D. Bosch, and W. J. Stiekema, "Adaptation of Spodoptera exigua larvae to plant proteinase inhibitors by induction of gut proteinase activity insensitive to inhibition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 17, pp. 8041–8045, 1995.

[32] M. C. Goulet, C. Dallaire, L. P. Vaillancourt et al., "Tailoring the specificity of a plant cystatin toward herbivorous insect digestive cysteine proteases by single mutations at positively selected amino acid sites," *Plant Physiology*, vol. 146, no. 3, pp. 1010–1019, 2008.

[33] K. Zhu-Salzman, H. Koiwa, R. A. Salzman, R. E. Shade, and J. E. Ahn, "Cowpea bruchid Callosobruchus maculatus uses a three-component strategy to overcome a plant defensive cysteine protease inhibitor," *Insect Molecular Biology*, vol. 12, no. 2, pp. 135–145, 2003.

[34] M. Volpicella, L. R. Ceci, J. Cordewener et al., "Properties of purified gut trypsin from Helicoverpa zea, adapted to proteinase inhibitors," *European Journal of Biochemistry*, vol. 270, no. 1, pp. 10–19, 2003.

[35] W. G. T. Willats, "Phage display: practicalities and prospects," *Plant Molecular Biology*, vol. 50, no. 6, pp. 837–854, 2002.

[36] W. Markland, A. C. Ley, S. W. Lee, and R. C. Ladner, "Iterative optimization of high-affinity protease inhibitors using phage

display. 1. Plasmin," *Biochemistry*, vol. 35, no. 24, pp. 8045–8057, 1996.

[37] M. Volpicella, C. Leoni, F. Arnesano, R. Gallerani, and L. R. Ceci, "Analysis by phage display selection and site-directed retromutagenesis of the Mustard Trypsin Inhibitor 2 reactive site," *Journal of Plant Physiology*, vol. 167, no. 17, pp. 1507–1511, 2010.

[38] F. R. Melo, M. O. Mello, O. L. Franco et al., "Use of phage display to select novel cystatins specific for Acanthoscelides obtectus cysteine proteinases," *Biochimica et Biophysica Acta—Proteins and Proteomics*, vol. 1651, no. 1-2, pp. 146–152, 2003.

[39] A. Hamdaoui, S. Wataleb, B. Devreese et al., "Purification and characterization of a group of five novel peptide serine protease inhibitors from ovaries of the desert locust, Schistocerca gregaria," *The FEBS Letters*, vol. 422, no. 1, pp. 74–78, 1998.

[40] H. Habib and K. M. Fazili, "Plant protease inhibitors: a defense strategy in plants," *Biotechnology and Molecular Biology Review*, vol. 2, no. 3, pp. 68–85, 2007.

[41] L. R. Ceci, M. Volpicella, Y. Rahbé, R. Gallerani, J. Beekwilder, and M. A. Jongsma, "Selection by phage display of a variant mustard trypsin inhibitor toxic against aphids," *The Plant Journal*, vol. 33, no. 3, pp. 557–566, 2003.

[42] M. A. Jongsma, P. L. Bakker, W. J. Stiekema, and D. Bosch, "Phage display of a double-headed proteinase inhibitor: analysis of the binding domains of potato proteinase inhibitor II," *Molecular Breeding*, vol. 1, no. 2, pp. 181–191, 1995.

[43] M. E. Santamaria, I. Cambra, M. Martinez et al., "Gene pyramiding of peptidase inhibitors enhances plant resistance to the spider mite Tetranychus urticae," *Public Library of Science ONE*, vol. 7, no. 8, article e43011, 2012.

[44] M. Chen, A. Shelton, and G. Y. Ye, "Insect-resistant genetically modified rice in china: from research to commercialization," *Annual Review of Entomology*, vol. 56, pp. 81–101, 2011.

[45] J. X. Huang, S. L. Bishop-Hurley, and M. A. Cooper, "Development of anti-infectives using phage display: biological agents against bacteria, viruses, and parasites," *Antimicrobial Agents and Chemotherapy*, vol. 56, no. 9, pp. 4569–4582, 2012.

[46] B. Liu, J. K. Hibbard, P. E. Urwin, and H. J. Atkinson, "The production of synthetic chemodisruptive peptides in planta disrupts the establishment of cyst nematodes," *Plant Biotechnology Journal*, vol. 3, no. 5, pp. 487–496, 2005.

[47] M. D. Winter, M. J. McPherson, and H. J. Atkinson, "Neuronal uptake of pesticides disrupts chemosensory cells of nematodes," *Parasitology*, vol. 125, no. 6, pp. 561–565, 2002.

[48] S. L. Bishop-Hurley, S. A. Mounter, J. Laskey et al., "Phage-displayed peptides as developmental agonists for Phytophthora capsici zoospores," *Applied and Environmental Microbiology*, vol. 68, no. 7, pp. 3315–3320, 2002.

[49] Z. D. Fang, J. G. Laskey, S. Huang et al., "Combinatorially selected defense peptides protect plant roots from pathogen infection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 49, pp. 18444–18449, 2006.

[50] C. K. Heng, S. M. Noor, T. S. Yee, and R. Y. Othman, "Biopanning for banana streak virus binding peptide by Phage display peptide library," *Journal of Biological Sciences*, vol. 7, no. 8, pp. 1382–1387, 2007.

[51] M. R. Safarnejad, G. S. Jouzani, M. Tabatabaei, R. M. Twyman, and S. Schillberg, "Antibody-mediated resistance against plant pathogens," *Biotechnology Advances*, vol. 29, no. 6, pp. 961–971, 2011.

[52] G. C. Whiteiam and W. Cockburn, "Antibody expression in transgenic plants," *Trends in Plant Science*, vol. 1, no. 8, pp. 268–272, 1996.

[53] K. C. Gough, W. Cockburn, and G. C. Whitelam, "Selection of phage-display peptides that bind to cucumber mosaic virus coat protein," *Journal of Virological Methods*, vol. 79, no. 2, pp. 169–180, 1999.

[54] F. W. Bai, H. W. Zhang, J. Yan et al., "Selection of phage-display peptides that bind specifically to the outer coat protein of Rice black streaked dwarf virus," *Acta Virologica*, vol. 46, no. 2, pp. 85–90, 2002.

[55] K. K. Pasumarthy, S. K. Mukherjee, and N. R. Choudhury, "The presence of tomato leaf curl Kerala virus AC3 protein enhances viral DNA replication and modulates virus induced gene-silencing mechanism in tomato plants," *Virology Journal*, vol. 8, article 178, 2011.

[56] S. H. Spoel and X. Dong, "How do plants achieve immunity? Defence without specialized immune cells," *Nature Reviews Immunology*, vol. 12, no. 2, pp. 89–100, 2012.

[57] T. Boller and G. Felix, "A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors," *Annual Review of Plant Biology*, vol. 60, pp. 379–407, 2009.

[58] S. T. Chisholm, G. Coaker, B. Day, and B. J. Staskawicz, "Host-microbe interactions: shaping the evolution of the plant immune response," *Cell*, vol. 124, no. 4, pp. 803–814, 2006.

[59] J. D. G. Jones and J. L. Dangl, "The plant immune system," *Nature*, vol. 444, no. 7117, pp. 323–329, 2006.

[60] C. Rioja, S. C. van Wees, K. A. Charlton, C. M. Pieterse, O. Lorenzo, and S. Garcia-Sanchez, "Wide screening of phage-displayed libraries identifies immune targets in planta," *Public Library of Science ONE*, vol. 8, no. 1, article e54654, 2013.

[61] A. Ma, Q. Hu, Z. Bai, Y. Qu, W. Liu, and G. Zhuang, "Functional display of fungal cellulases from Trichoderma reesei on phage M13," *World Journal of Microbiology and Biotechnology*, vol. 24, no. 10, pp. 2003–2009, 2008.

[62] T. Beliën, K. Hertveldt, K. van den Brande, J. Robben, S. van Campenhout, and G. Volckaert, "Functional display of family 11 endoxylanases on the surface of phage M13," *Journal of Biotechnology*, vol. 115, no. 3, pp. 249–260, 2005.

[63] T. Serizawa, K. Iida, H. Matsuno, and K. Kurita, "Cellulose-binding heptapeptides identified by phage display methods," *Chemistry Letters*, vol. 36, no. 8, pp. 988–989, 2007.

[64] L. C. Gunnarsson, Q. Zhou, C. Montanier, E. N. Karlsson, H. Brumer, and M. Ohlin, "Engineered xyloglucan specificity in a carbohydrate-binding module," *Glycobiology*, vol. 16, no. 12, pp. 1171–1180, 2006.

[65] L. C. Gunnarsson, E. N. Karlsson, A. S. Albrekt, M. Andersson, O. Holst, and M. Ohlin, "A carbohydrate binding module as a diversity-carrying scaffold," *Protein Engineering, Design and Selection*, vol. 17, no. 3, pp. 213–221, 2004.

[66] L. C. Gunnarsson, C. Montanier, R. B. Tunnicliffe et al., "Novel xylan-binding properties of an engineered family 4 carbohydrate-binding module," *Biochemical Journal*, vol. 406, no. 2, pp. 209–214, 2007.

[67] T. Beliën, S. van Campenhout, A. Vanden Bosch et al., "Engineering molecular recognition of endoxylanase enzymes and their inhibitors through phage display," *Journal of Molecular Recognition*, vol. 20, no. 2, pp. 103–112, 2007.

[68] T. M. Bourgois, D. V. Nguyen, S. Sansen et al., "Targeted molecular engineering of a family 11 endoxylanase to decrease its

sensitivity towards Triticum aestivum endoxylanase inhibitor types," *Journal of Biotechnology*, vol. 130, no. 1, pp. 95–105, 2007.

[69] T. Beliën, I. J. Joye, J. A. Delcour, and C. M. Courtin, "Computational design-based molecular engineering of the glycosyl hydrolase family 11 B. subtilis XynA endoxylanase improves its acid stability," *Protein Engineering, Design & Selection*, vol. 22, no. 10, pp. 587–596, 2009.

[70] C. Hadley, "Food allergies on the rise? Determining the prevalence of food allergies, and how quickly it is increasing, is the first step in tackling the problem," *The EMBO Reports*, vol. 7, no. 11, pp. 1080–1083, 2006.

[71] R. M. Helm, "Allergy to plant seed proteins," *Journal of New Seeds*, vol. 3, no. 3, pp. 37–60, 2001.

[72] A. K. Verma, S. Kumar, M. Das, and P. D. Dwivedi, "A comprehensive review of legume allergy," *Clinical Reviews in Allergy & Immunology*, 2013.

[73] C. Rhyner, M. Weichel, S. Flückiger, S. Hemmann, T. Kleber-Janke, and R. Crameri, "Cloning allergens via phage display," *Methods*, vol. 32, no. 3, pp. 212–218, 2004.

[74] T. Kleber-Janke, R. Crameri, S. Scheurer, S. Vieths, and W. M. Becker, "Patient-tailored cloning of allergens by phage display: peanut (Arachis hypogaea) profilin, a food allergen derived from a rare mRNA," *Journal of Chromatography B: Biomedical Sciences and Applications*, vol. 756, no. 1-2, pp. 295–305, 2001.

[75] J. M. Davies, R. E. O'Hehir, and C. Suphioglu, "Use of phage display technology to investigate allergen-antibody interactions," *Journal of Allergy and Clinical Immunology*, vol. 105, no. 6, pp. 1085–1092, 2000.

[76] S. H. Sicherer, A. Muñoz-Furlong, and H. A. Sampson, "Prevalence of peanut and tree nut allergy in the United States determined by means of a random digit dial telephone survey: a 5-year follow-up study," *Journal of Allergy and Clinical Immunology*, vol. 112, no. 6, pp. 1203–1207, 2003.

[77] J. Grundy, S. Matthews, B. Bateman, T. Dean, and S. H. Arshad, "Rising prevalence of allergy to peanut in children: data from 2 sequential cohorts," *Journal of Allergy and Clinical Immunology*, vol. 110, no. 5, pp. 784–789, 2002.

[78] A. T. Clark and P. W. Ewan, "Good prognosis, clinical features, and circumstances of peanut and tree nut reactions in children treated by a specialist allergy center," *Journal of Allergy and Clinical Immunology*, vol. 122, no. 2, pp. 286–289, 2008.

[79] T. Kleber-Janke, R. Crameri, U. Appenzeller, M. Schlaak, and W. M. Becker, "Selective cloning of peanut allergens, including profilin and 2S albumins, by phage display technology," *International Archives of Allergy and Immunology*, vol. 119, no. 4, pp. 265–274, 1999.

[80] M. P. de Leon, J. M. Rolland, and R. E. O'Hehir, "The peanut allergy epidemic: allergen molecular characterisation and prospects for specific therapy," *Expert Reviews in Molecular Medicine*, vol. 9, no. 1, pp. 1–18, 2007.

[81] C. Bittner, B. Grassau, K. Frenzel, and X. Baur, "Identification of wheat gliadins as an allergen family related to baker's asthma," *Journal of Allergy and Clinical Immunology*, vol. 121, no. 3, pp. 744–749, 2008.

[82] S. Mahadov and P. H. Green, "Celiac disease: a challenge for all physicians," *Gastroenterology & Hepatology*, vol. 7, no. 8, pp. 554–556, 2011.

[83] L. M. Sollid and B. Jabri, "Is celiac disease an autoimmune disorder?" *Current Opinion in Immunology*, vol. 17, no. 6, pp. 595–600, 2005.

[84] W. Dieterich, T. Ehnis, M. Bauer et al., "Identification of tissue transglutaminase as the autoantigen of celiac disease," *Nature Medicine*, vol. 3, no. 7, pp. 797–801, 1997.

[85] Ø. Molberg, S. N. McAdam, and L. M. Sollid, "Role of tissue transglutaminase in celiac disease," *Journal of Pediatric Gastroenterology and Nutrition*, vol. 30, no. 3, pp. 232–240, 2000.

[86] J. E. Bernardin, D. D. Kasarda, and D. K. Mecham, "Preparation and characterization of alpha-gliadin," *Journal of Biological Chemistry*, vol. 242, no. 3, pp. 445–450, 1967.

[87] W. Dieterich, B. Esslinger, and D. Schuppan, "Pathomechanisms in celiac disease," *International Archives of Allergy and Immunology*, vol. 132, no. 2, pp. 98–108, 2003.

[88] K. Hoffmann, M. Alminger, T. Andlid, T. Chen, O. Olsson, and A. S. Sandberg, "Blocking peptides decrease tissue transglutaminase processing of gliadin *in vitro*," *Journal of Agricultural and Food Chemistry*, vol. 57, no. 21, pp. 10150–10155, 2009.

[89] T. Chen, K. Hoffmann, S. Östman, A. S. Sandberg, and O. Olsson, "Identification of gliadin-binding peptides by phage display," *BMC Biotechnology*, vol. 11, no. 16, 2011.

[90] E. Jensen-Jarolim, A. Leitner, H. Kalchhauser et al., "Peptide mimotopes displayed by phage inhibit antibody binding to Bet v 1, the major birch pollen allergen, and induce specific IgG response in mice," *The FASEB Journal*, vol. 12, no. 15, pp. 1635–1642, 1998.

[91] T. Mothes, "Deamidated gliadin peptides as targets for celiac disease-specific antibodies," *Advances in Clinical Chemistry*, vol. 44, pp. 35–63, 2007.

[92] M. B. Mudgett and S. Clarke, "Hormonal and environmental responsiveness of a developmentally regulated protein repair L-isoaspartyl methyltransferase in wheat," *Journal of Biological Chemistry*, vol. 269, no. 41, pp. 25605–25612, 1994.

[93] C. Job, L. Rajjou, Y. Lovigny, M. Belghazi, and D. Job, "Patterns of protein oxidation in Arabidopsis seeds and during germination," *Plant Physiology*, vol. 138, no. 2, pp. 790–802, 2005.

[94] K. P. Lu, G. Finn, T. H. Lee, and L. K. Nicholson, "Prolyl cis-trans isomerization as a molecular timer," *Nature Chemical Biology*, vol. 3, no. 10, pp. 619–629, 2007.

[95] J. Sanchez, B. J. Nikolau, and P. K. Stumpf, "Reduction of N-acetyl methionine sulfoxide in plants," *Plant Physiology*, vol. 73, no. 3, pp. 619–623, 1983.

[96] U. M. N. Murthy and W. Q. Sun, "Protein modification by Amadori and Maillard reactions during seed storage: roles of sugar hydrolysis and lipid peroxidation," *Journal of Experimental Botany*, vol. 51, no. 348, pp. 1221–1228, 2000.

[97] M. B. Mudgett and S. Clarke, "Characterization of plant L-isoaspartyl methyltransferases that may be involved in seed survival: purification, cloning, and sequence analysis of the wheat germ enzyme," *Biochemistry*, vol. 32, no. 41, pp. 11100–11111, 1993.

[98] E. Chatelain, P. Satour, E. Laugier et al., "Evidence for participation of the methionine sulfoxide reductase repair system in plant seed longevity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 9, pp. 3633–3638, 2013.

[99] P. Verma, A. Singh, H. Kaur, and M. Majee, "Protein L-isoaspartyl methyltransferase1 (CaPIMT1) from chickpea mitigates oxidative stress-induced growth inhibition of Escherichia coli," *Planta*, vol. 231, no. 2, pp. 329–336, 2010.

[100] K. Oracz, H. E. M. Bouteau, J. M. Farrant et al., "ROS production and protein oxidation as a novel mechanism for seed dormancy alleviation," *The Plant Journal*, vol. 50, no. 3, pp. 452–465, 2007.

[101] H. A. Doyle, R. J. Gee, and M. J. Mamula, "Altered immunogenicity of isoaspartate containing proteins," *Autoimmunity*, vol. 40, no. 2, pp. 131–137, 2007.

[102] D. W. Aswad, M. V. Paranandi, and B. T. Schurter, "Isoaspartate in peptides and proteins: formation, significance, and analysis," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 21, no. 6, pp. 1129–1136, 2000.

[103] D. Ingrosso, A. F. Perna, S. D'Angelo et al., "Enzymatic detection of L-isoaspartyl residues in food proteins and the protective properties of trehalose," *Journal of Nutritional Biochemistry*, vol. 8, no. 9, pp. 535–540, 1997.

[104] L. Ogé, G. Bourdais, J. Bove et al., "Protein repair L-Isoaspartyl methyltransferasel is involved in both seed longevity and germination vigor in arabidopsis," *The Plant Cell*, vol. 20, no. 11, pp. 3022–3037, 2008.

[105] P. Verma, H. Kaur, B. P. Petla, V. Rao, S. C. Saxena, and M. Majee, "PROTEIN L-ISOASPARTYL METHYLTRANSFERASE2 is differentially expressed in chickpea and enhances seed vigor and longevity by reducing abnormal isoaspartyl accumulation predominantly in seed nuclear proteins," *Plant Physiology*, vol. 161, no. 3, pp. 1141–1157, 2013.

[106] C. W. Vertucci, "Effects of cooling rate on seeds exposed to liquid nitrogen temperatures," *Plant Physiology*, vol. 90, no. 4, pp. 1478–1485, 1989.

[107] R. H. Ellis, T. D. Hong, and E. H. Roberts, "A low-moisture-content limit to logarithmic relations between seed moisture content and longevity," *Annals of Botany*, vol. 61, no. 4, pp. 405–408, 1988.

[108] H. A. Doyle, R. J. Gee, and M. J. Mamula, "A failure to repair self-proteins leads to T cell hyperproliferation and autoantibody production," *Journal of Immunology*, vol. 171, no. 6, pp. 2840–2847, 2003.

[109] K. J. Reissner and D. W. Aswad, "Deamidation and isoaspartate formation in proteins: unwanted alterations or surreptitious signals?" *Cellular and Molecular Life Sciences*, vol. 60, no. 7, pp. 1281–1295, 2003.

[110] J. Lanthier and R. R. Desrosiers, "Protein L-isoaspartyl methyltransferase repairs abnormal aspartyl residues accumulated *in vivo* in type-I collagen and restores cell migration," *Experimental Cell Research*, vol. 293, no. 1, pp. 96–105, 2004.

[111] R. Kern, A. Malki, J. Abdallah et al., "Protein isoaspartate methyltransferase is a multicopy suppressor of protein aggregation in Escherichia coli," *Journal of Bacteriology*, vol. 187, no. 4, pp. 1377–1383, 2005.

[112] Q. Xu, M. P. Belcastro, S. T. Villa, R. D. Dinkins, S. G. Clarke, and A. B. Downie, "A second protein L-isoaspartyl methyltransferase gene in Arabidopsis produces two transcripts whose products are sequestered in the nucleus," *Plant Physiology*, vol. 136, no. 1, pp. 2652–2664, 2004.

[113] S. T. Villa, Q. Xu, A. B. Downie, and S. G. Clarke, "Arabidopsis protein repair L-isoaspartyl methyltransferases: predominant activities at lethal temperatures," *Physiologia Plantarum*, vol. 128, no. 4, pp. 581–592, 2006.

[114] V. Vigneswara, J. D. Lowenson, C. D. Powell et al., "Proteomic identification of novel substrates of a protein isoaspartyl methyltransferase repair enzyme," *Journal of Biological Chemistry*, vol. 281, no. 43, pp. 32619–32629, 2006.

[115] J. X. Zhu, H. A. Doyle, M. J. Mamula, and D. W. Aswad, "Protein repair in the brain, proteomic analysis of endogenous substrates for protein L-isoaspartyl methyltransferase in mouse brain," *Journal of Biological Chemistry*, vol. 281, no. 44, pp. 33802–33813, 2006.

[116] R. D. Dinkins, S. M. Majee, N. R. Nayak et al., "Changing transcriptional initiation sites and alternative 5'- and 3'-splice site selection of the first intron deploys Arabidopsis protein isoaspartyl methyltransferase2 variants to different subcellular compartments," *The Plant Journal*, vol. 55, no. 1, pp. 1–13, 2008.

[117] T. Chen, N. Nayak, S. M. Majee et al., "Substrates of the Arabidopsis thaliana protein isoaspartyl methyltransferase 1 identified using phage display and biopanning," *Journal of Biological Chemistry*, vol. 285, no. 48, pp. 37281–37292, 2010.

[118] M. E. Hudson, T. Bruggink, S. H. Chang et al., "Analysis of gene expression during Brassica seed germination using a cross-species microarray platform," *Crop Science*, vol. 47, no. 2, pp. S96–S112, 2007.

[119] T. W. O'Brien, "Evolution of a protein-rich mitochondrial ribosome: implications for human genetic disease," *Gene*, vol. 286, no. 1, pp. 73–79, 2002.

[120] E. H. Harris, J. E. Boynton, and N. W. Gillham, "Chloroplast ribosomes and protein synthesis," *Microbiological Reviews*, vol. 58, no. 4, pp. 700–754, 1994.

[121] L. Dure III, S. C. Greenway, and G. A. Galau, "Developmental biochemistry of cottonseed embryogenesis and germination: changing messenger ribonucleic acid populations as shown by *in vitro* and *in vivo* protein synthesis," *Biochemistry*, vol. 20, no. 14, pp. 4162–4168, 1981.

[122] G. A. Galau, D. W. Hughes, and L. I. Dure, "Developmental biochemistry of cottonseed embryogenesis and germination: changing messenger ribonucleic acid populations as shown by reciprocal heterologous complementary deoxyribonucleic acid-messenger ribonucleic acid hybridizationembryogenesis-abundant (LEA) mRNAs," *Plant Molecular Biology*, vol. 7, no. 3, pp. 155–170, 1986.

[123] V. Mattimore and J. R. Battista, "Radioresistance of Deinococcus radiodurans: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation," *Journal of Bacteriology*, vol. 178, no. 3, pp. 633–637, 1996.

[124] D. Billi and M. Potts, "Life and death of dried prokaryotes," *Research in Microbiology*, vol. 153, no. 1, pp. 7–12, 2002.

[125] I. Kranner, W. J. Cram, M. Zorn et al., "Antioxidants and photoprotection in a lichen as compared with its isolated symbiotic partners," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 8, pp. 3141–3146, 2005.

[126] S. Hengherr, A. G. Heyer, H. R. Köhler, and R. O. Schill, "Trehalose and anhydrobiosis in tardigrades—evidence for divergence in responses to dehydration," *The FEBS Journal*, vol. 275, no. 2, pp. 281–288, 2008.

[127] A. Tunnacliffe and J. Lapinski, "Resurrecting Van Leeuwenhoek's rotifers: a reappraisal of the role of disaccharides in anhydrobiosis," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1438, pp. 1755–1771, 2003.

[128] J. S. Clegg, "Cryptobiosis—a peculiar state of biological organization," *Comparative Biochemistry and Physiology—B Biochemistry and Molecular Biology*, vol. 128, no. 4, pp. 613–624, 2001.

[129] J. H. Crowe, F. A. Hoekstra, and L. M. Crowe, "Anhydrobiosis," *Annual Review of Physiology*, vol. 54, pp. 579–599, 1992.

[130] J. Browne, A. Tunnacliffe, and A. Burnell, "Anhydrobiosis: plant desiccation gene found in a nematode," *Nature*, vol. 416, no. 6876, p. 38, 2002.

[131] J. A. Browne, K. M. Dolan, T. Tyson, K. Goyal, A. Tunnacliffe, and A. M. Burnell, "Dehydration-specific induction of hydrophilic protein genes in the anhydrobiotic nematode

Aphelenchus avenae," *Eukaryotic Cell*, vol. 3, no. 4, pp. 966–975, 2004.

[132] M. J. Oliver, Z. Tuba, and B. D. Mishler, "The evolution of vegetative desiccation tolerance in land plants," *Plant Ecology*, vol. 151, no. 1, pp. 85–100, 2000.

[133] E. H. Muslin and P. H. Homann, "Light as a hazard for the desiccation-resistant "resurrection" fern Polypodium polypodioides L," *Plant, Cell & Environment*, vol. 15, no. 1, pp. 81–89, 1992.

[134] T. S. Stuart, "Revival of respiration and photosynthesis in dried leaves of Polypodium polypodioides," *Planta*, vol. 83, no. 2, pp. 185–206, 1968.

[135] J. P. Moore, N. T. Le, W. F. Brandt, A. Driouich, and J. M. Farrant, "Towards a systems-based understanding of plant desiccation tolerance," *Trends in Plant Science*, vol. 14, no. 2, pp. 110–117, 2009.

[136] O. Leprince and J. Buitink, "Desiccation tolerance: from genomics to the field," *Plant Science*, vol. 179, no. 6, pp. 554–564, 2010.

[137] G. Barker, *The Agricultural Revolution in Prehistory: Why Did Foragers Become Farmers?* Oxford University Press, Oxford, UK, 2006.

[138] D. Z. Li and H. W. Pritchard, "The science and economics of *ex situ* plant conservation," *Trends in Plant Science*, vol. 14, no. 11, pp. 614–621, 2009.

[139] M. Hundertmark, J. Buitink, O. Leprince, and D. K. Hincha, "The reduction of seed-specific dehydrins reduces seed longevity in Arabidopsis thaliana," *Seed Science Research*, vol. 21, no. 3, pp. 165–173, 2011.

[140] T. W. Hegarty, "The physiology of seed hydration and dehydration, and the relation between water stress and the control of germination: a review," *Plant, Cell and Environment*, vol. 1, no. 2, pp. 101–119, 1978.

[141] J. Buitink, B. L. Vu, P. Satour, and O. Leprince, "The re-establishment of desiccation tolerance in germinated radicles of Medicago truncatula Gaertn. Seeds," *Seed Science Research*, vol. 13, no. 4, pp. 273–286, 2003.

[142] J. Maia, B. J. Dekkers, N. J. Provart, W. Ligterink, and H. W. Hilhorst, "The re-establishment of desiccation tolerance in germinated Arabidopsis thaliana seeds and its associated transcriptome," *Public Library of Science ONE*, vol. 6, no. 12, article e29123, 2011.

[143] M. J. Wise and A. Tunnacliffe, "POPP the question: what do LEA proteins do?" *Trends in Plant Science*, vol. 9, no. 1, pp. 13–17, 2004.

[144] A. Tunnacliffe and M. J. Wise, "The continuing conundrum of the LEA proteins," *Naturwissenschaften*, vol. 94, no. 10, pp. 791–812, 2007.

[145] A. Tunnacliffe, D. K. Hincha, O. Leprince, and D. Macherel, "LEA proteins: versatility of form and function," in *Dormancy and Resistance in Harsh Environments*, E. Lubzens, J. Cerdáa, and M. S. Clark, Eds., pp. 91–108, Springer, Berlin, Germany, 2010.

[146] R. Kushwaha, T. D. Lloyd, K. R. Schäfermeyer, S. Kumar, and A. B. Downie, "Identification of late embryogenesis abundant (LEA) protein putative interactors using phage display," *International Journal of Molecular Sciences*, vol. 13, no. 1, pp. 6582–6603, 2012.