Boffin Access

Open Access Journal | Volume: 1.1

Systems and Computational Biology Journal

Pseudorandom Vs. Random Polymers - How to Improve the Efficiency of Lithography-based Synthesis

Phillip Stafford* | VP Bioinformatics, Caris Life Sciences, Phoenix, Arizona, USA

Abstract

Computer lithography has been used for decades in the electronics industry. Biochemists have used light-directed lithography and photoacids and photobases to enable BOC- and FMOC-based peptide synthesis, enabling the creation of high-density peptide microarrays. By reducing the size of the features, more peptides can be produced on industry-standard microscope slides. Phage display has set the standard for the utility of probing a high density peptide library. However, a sufficiently high-density peptide microarray can produce similar results in a few hours *vs.* weeks of panning and selection.

For the specific case of *random* peptide microarrays, we examined a method to reduce manufacturing costs by exchanging perfectly random peptides to pseudorandom ones. The simplest mathematical formula for predicting the number of synthesis steps is M (steps)=AA (# of different amino acids) * L (peptide length). However, when M< AA*L, peptides can no longer be perfectly random, and a sequence bias will result. M affects synthesis time, AA and L affect performance given that performance is directly related to the lack of sequence bias. We isolated M as a variable and tested how different values of M affect the performance of immunosignatures. Immunosignatures require random peptides to detect patterns of antibody binding. As M decreases, the peptides gain sequence bias. The immunosignature assay is used to quantify the impact of M. We first measured the predicted properties of peptide libraries for M=35, 70, 140 and 272, then we synthesized actual peptides and tested how 51 different monoclonal antibodies bound to the non-random peptides. We then tested whether *Coccidiomycoses possadasii* (Valley Fever) infections could be distinguished – a very simple task for a random peptide library. At low M the disorder in the peptide library decreases, and one sees a near-monotonic signal with low information content. There is a value of M, however, where synthesis steps can be reduced without unduly compromising the information content of the resulting immunosignature. This tunable parameter should be examined as a way to reduce synthesis time and manufacturing costs for random peptide microarrays.

Keywords

in situ peptide synthesis; Photolithography; Random peptide; Immunosignature

Introduction

Many peptide microarrays are in the form of tiled epitope arrays [1]. They are commonly used to map antibodies to a segment of a protein. However, there are other uses for peptide microarrays-cell binding [2], small-molecule binding [3], protein binding [4], kinase activity [5], even enzyme modulation [6] and a process that we use called 'immunosignaturing' [7]. Epitope arrays rely on antibodies recognizing and binding the exact sequence against which they were raised. Immunosignatures, on the other hand, use random sequence peptides and do not rely on perfect antibody-peptide matches. On an immunosignature peptide microarray, antibodies can bind to their cognate peptide, but they can also bind apparently unrelated peptides with equal or higher affinity [8,9]. This non-cognate binding may be to mimotopes of the eliciting antigen [10]. Another explanation is that off-target binding is an intrinsic property of an antibody, theoretically as important as on-target binding. Immunosignatures detect both properties. There is a great deal of information that can be extracted from an immunosignature assay, even to the point of diagnosing disease [11-14].

Epitope arrays can be very expensive to manufacture, limiting their use to smaller experiments. Immunosignature arrays can be produced at very high numbers, since they use manufacturing processes taken from the semiconductor industry. Increasing the number of arrays, manufactured would actually reduce the cost per unit. We will describe a method to reduce manufacturing complexity and describe how simplifying manufacturing affects the randomness of a peptide library and how randomness affects immunosignatures.

Article Information

DOI: 10.31021/scbj.20181102
Article Type: Research Article
Journal Type: Open Access
Volume: 1 **Issue:** 1
Manuscript ID: SCBJ-1-102
Publisher: Boffin Access Limited

Received Date: 07 December 2018
Accepted Date: 27 December 2018
Published Date: 29 December, 2018

*Corresponding author:

Phillip Stafford
VP Bioinformatics
Caris Life Sciences, Phoenix
Arizona, USA
E-mail: pstafford@carisls.com

Citation: Stafford P. Pseudorandom Vs. Random Peptides-How to Improve the Efficiency of Lithography-based Synthesis of Peptide Microarrays. Syst Comput Biol J. 2018 Dec;1(1):102

Immunosignatures use random sequence peptides to gauge specificity of antibodies

Immunosignatures measure the binding between random peptides and antibodies. Immunosignatures measure both on-target and off-target binding of antibodies simultaneously [8]. As a reference, phage display or phage panning probe extremely large libraries. The selection process leaves only strong binders. Those survivors are usually closely related to the actual epitope. In contrast, peptide microarrays do not utilize a selection process. All of the binding data is retained, and is scalable. Phage libraries can be extremely large [15] which is to their benefit, while microarrays typically contain a few million peptides at best [16]. However, each peptide interaction can be more informative than the final pool of phage display peptides. In addition, even a relatively small random peptide library contains many unique epitopes [17-19] if longer peptides are used [1,14], while phage display often relies on 6-10mers [20]. Phage libraries may be diverse but microarrays can extract low binders as well as high binders. By analyzing binding intensities one can identify residues at key positions that play a role in enhancing binding *vs.* those that prevent binding. When sera or an antibody is applied to a random peptide microarray, the resulting 'immunosignature' [7] can provide information about how many different peptides the antibody binds, how diverse are the binding sequences, where the critical residues lie in the epitope, and how amino acids bind to the paratope. The immunosignature can be so informative and reproducible that diseases can be diagnosed with high accuracy [11,13,14,21,22]. This diagnostic feature does not *require* information about peptide sequences but even so, the peptide sequences in a signature can still predict linear epitopes from infectious agents [17] or cancer antigens [19].

Rationale and overview of concept

Immunosignatures have the potential to diagnose disease [11-14,16,23] and monitor health [24,25]. This is not really practical unless the microarrays can be made cheaply, reproducibly and in high number. We suggest that reducing the synthesis time for the arrays will reduce the cost, reduce the potential for mistakes, and improve reproducibility. The coupling step required to build peptides from amino acids can be shortened to an extent, but the fidelity of solid state synthesis is directly correlated to coupling time [26]. In the case where many peptides are synthesized simultaneously by 006 Cithography, and where random peptides are required, synthesis can be shortened by changing the number of steps. This works for any method: light directed (ex. Nimblegen), electrochemistry (ex. Combimatrix), or dry-chemistry (PEPper Print). In this study we used masks to create features which are exposed to UV light which, upon illuminating a photoacid generator (PAG), creates a local region of high acid concentration removing the protecting group and allowing the next amino acid to couple. Peptides require exactly AA multiplied by L steps. AA represents the total number of different amino acids. L is the length of the resulting peptide. Any number of synthesis steps less than AA*L reduces the randomness of the resulting peptides, creating a bias. Pseudorandom peptides therefore require <AA*L steps. We present data to examine how pseudorandom sequences affect immunosignatures, the concept applies to any chemical synthesis where random or pseudorandom libraries are created. We will investigate performance characteristics of an immunosignature library by changing from random to pseudorandom sequences. Pseudorandom peptides can be made faster than true random peptides. Our hypothesis is that a bias introduced to a random peptide library will change detection sensitivity. However, it is possible that immunosignatures do not require unbiased sequences to work. It is possible that a pseudorandom peptide library still retains the important characteristics of a random library. The original immunosignature microarray contained 10,000 peptides [7] which were spotted by piezo printing. It was decided to expand the library size by at least 10-fold to meet future needs; the newest immunosignature arrays have 330,000 peptides [16]. The only method that could produce this size library was *in situ* peptide synthesis.

In situ peptide synthesis

There are many ways to manufacture peptide microarrays. For a review see Gao et al. [27,28]. Light-directed chemical synthesis has been quite successful, dating back to 1995 (Fodor et al. [29], US Patent 5405783A). Some synthesis methods, direct light on a photoacid [16] or photobase [30] generator which removes amino acid protecting groups. Other methods use photoactivatable amino acids [29]. Some methods use shadow masks [16], some use digital micromirrors [31], some utilize a peptide 'toner' and a 20 'color' laser printer [32], while still others use electrical circuits to alter the pH in a microwell [33]. Each of these synthesis methods relies on the systematic addition of amino acids to a growing peptide. For random peptide libraries, all of these methods can take advantage of reduced synthetic steps using pseudorandom peptides.

One of the highest throughput methods for the production of peptide microarrays is shadow-mask lithography [34]. Shadow-masks are pre-made so they are not as flexible as programmable light-directed arrays, and they require substantial up-front costs. However, by implementing high-resolution steppers, automated aligners, synthesizers and large wafers, mask-based peptide microarrays can be very inexpensive, high density, high precision and high throughput. These characteristics are important for high volume/low costproduction.

Random *vs.* pseudorandom libraries

Many immunological experiments make use of random sequence peptides [35]. Phage display uses random sequence libraries [36] and has been used to identify therapeutics [37], epitopes [38], even enzyme modulators [39]. Panning large libraries can identify very specific biomarkers, but these methods are neither cheap nor high throughput. Immunosignatures use a fixed library of a given size for any sample. A diverse surface of 3D shapes produced by random sequence peptides has been very successful in many applications [11,13,21,40-42]. Surprisingly, 10,000 peptides were sufficient to distinguish over 15 different diseases simultaneously, with >95% accuracy [14]. However, 10,000 peptides likely had a finite level of specificity and sensitivity. It was decided to create at least 300,000 peptides for the next immunosignature iteration. When M=AA*L, synthesis of 300,000 17mers would take over 2 weeks of continuous synthesis. Altering the library from random to pseudorandom could drastically reduce manufacturing time.

How *in-situ* peptide synthesis is affected by amino acid selection

The simplest algorithm for the number of synthesis steps (or photomasks) required to produce a peptide of length L is M (number of masks)=AA (number of amino acids used) * L (length). For a 17mer using 20 amino acids, 20*17=340 synthetic steps are required. In mask-based peptide photolithography, 340 steps would take over 2 weeks to complete. For immunosignature peptides, M can be less than AA*L, but the resulting peptides become *pseudo*random. Each mask costs several thousand dollars. Each synthesis step takes approximately 15 minutes. It is beneficial to reduce M. Figure 1 illustrates the bias imposed by forcing M to be less than AA*L.

Experimental design

We asked two questions about reducing synthesis steps. First, how do the chemical characteristics of the peptide library change as mask numbers are reduced? Second, how does reduced diversity alter the outcome of a biochemical assay? The first question was addressed using predicted characteristics of hypothetical peptides. For the second question, an immunosignature of monoclonal antibodies and an immunosignature of human sera are used to measure performance. We first established the starting parameters. Kuznetsov [43] noted that some amino acids are preferentially used in immunosignatures [43]. To maintain consistency with the actual immunosignature arrays that will be manufactured, we eliminated Methionine, Threonine and Isoleucine because they appear rarely in peptides that make up diagnostic immunosignatures [43]. Cysteine

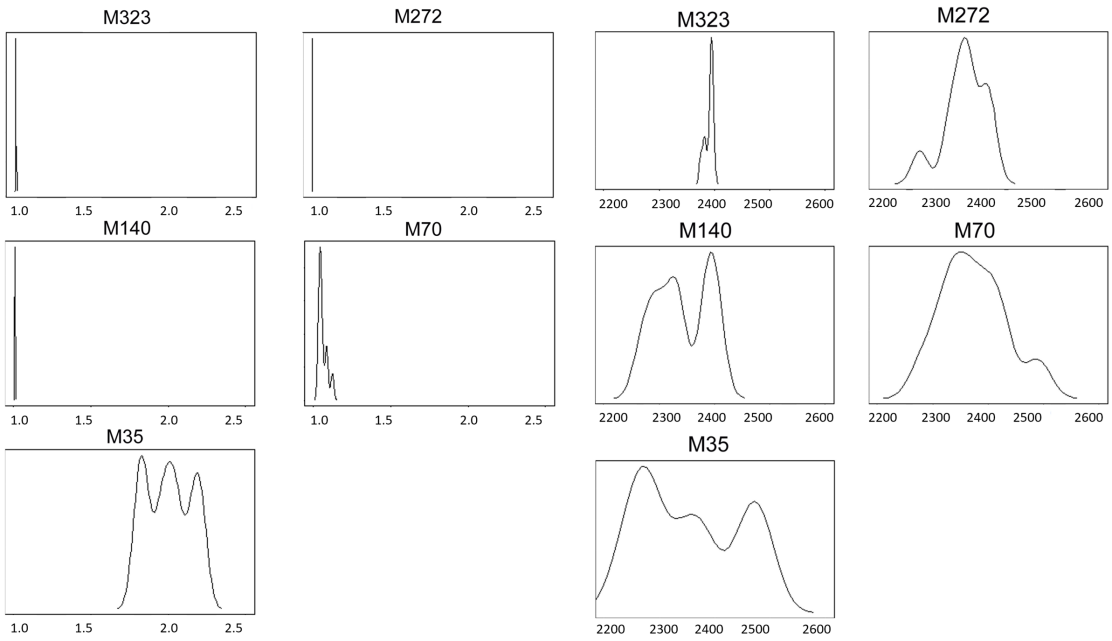

Figure 1: Analysis of amino acids prevalence across each peptide position using mask restrictions.

100,000 peptides were contructed *in silico* using a randomization method that takes into account the total number of shadow masks used in synthesis of 17mer peptides. Each graph shows the number of times each of the 16 amino acids is used at each of the 17 positions for a given number of masks. The X-axis for each panel represents the position in the peptide from C-terminus (position 1, facing the surface of the microarray) to the N-terminus (position 17, facing away from the surface). The Y-axis is the number of times each amino acid appears at that position. The legend associates graph colors with an amino acid. The lower middle graph is M_{323}, and is used as a baseline comparison. M_{323} is unique as it uses 19 amino acids (excluding cysteine) and mimics a library of random peptides if synthesized on beads and cleaved/purified, with no mask number restrictions. M_{323} should have no positional bias. M=35 (upper left) shows the highest positional bias, suggesting that diversity of sequence would be compromised using only 35 masks, even though the number of steps is nearly 10-fold less than fully random. M=70 (middle top), M=140 (right top), and M=272 (bottom left) show progressively decreasing bias trends, and smoother amino acid distributions. M=323 has the lowest variation in positional bias across the length of the peptide since each position was determined by a random number generator, independently.

can form disulfide bonds under aqueous conditions, and was removed for that reason. Thus our virtual 17mer peptide library used 16 amino acids (no C, T, M, or I). We created two control sets: M=272 uses 16 amino acids and no other mask restrictions. M=323 leaves out only Cysteine. The M=323 case matches exactly the existing 10,000 peptide spotted peptide library that has been used in several early publications (NCBI GEO accession GPL17600). We also created intermediate test sets where M=35, M=70, and M=140.

The first analysis was completed using 100,000 virtual peptides. For each library, an algorithm (see Supplemental Information) creates *M* masks with 100,000 features and 'synthesizes' peptides using these virtual masks. The resulting library is analyzed for chemical properties using ProtParam ((http://web.expasy.org/protparam/).

The second analysis involved synthesizing a library of actual peptides corresponding to the same mask restrictions. 384 peptides were selected at random from each 100,000 virtual peptide libraries and sent to Sigma Genosys (St. Louis, MO) for synthesis. We already had several M=323 10,000 peptide microarrays made, so the human sera experiments for M=323 were done by randomly picking 384 peptides from the 10,000 existing peptides to simulate a 384-peptide library. Each of these 384-peptide sets (except M=323) were tested using 51 different commercial monoclonals (Supplemental Information Table 1). We previously published data on monoclonals using the 10,000 peptide microarray [7]. Human sera from 8 controls (healthy) humans and 8 patients diagnosed with *Coccidiomycoses* [11] were used for all libraries including M=323.

Materials and Methods

In silico peptide generation

We created an R script called "Peptide Generator" in order to simulate how peptides would be created when the number of synthetic steps is less than AA*L (Supplemental Information). The program uses the following variables:

P is the library size, number of peptides

M is the number of virtual masks

N is peptide length

DD is the number of amino acids used

aa is the one-letter amino acid code

peptide is the output of *P* peptides

The script can use one-letter code for D- or L-amino acids. The algorithm currently uses loops to cycle through the synthesis steps to allow non-coders to clearly see the steps, but for efficiency the loops should be replaced by 'apply' statements.

To compare the predicted chemical characteristics of peptides made by Peptide Generator, we created 100,000 virtual 17mer peptides using 16 amino different acids for each mask restriction scenario (M=35, 70, 140 and 272) as well as 100,000 completely random peptides using 19 amino acids (M=323). The method for creating the peptide libraries is by the greedy algorithm, where masks are made available as amino acids are cycled through one by one. The amino acids are added one step at a time to the growing peptide chain. The program cycles through each amino acid alphabetically and stops when every peptide is exactly 17 amino acids long and every mask is used once. The algorithm as written does not attempt to make the distribution of amino acids even during synthesis. As a result, some peptides tend to catch up in length as the masks run out by quickly adding any available amino acid(s) to the N-term of the synthesis. This creates a bias in sequence,driven by the order of the amino acid addition. This bias was deemed acceptable since each peptide library has the same bias, and performance comparisons were relative. Methods to artificially compensate for this bias introduced an unpredictable distribution that was more complex to decipher.

M	Avg. CC Ab	Avg. CC serum	ANOVA p-val	Avg. raw value Ab	Median raw value Ab	CV
35	0.451	0.989	$p<4.00 \times 10^{-37}$	745	238	6%
70	0.480	0.971	$p<2.49 \times 10^{-43}$	1238	311	21%
140	0.361	0.959	$p<1.32 \times 10^{-41}$	1334	367	22%
272	0.358	0.944	$p<8.99 \times 10^{-56}$	1607	428	22%
323		0.861	$p<1.13 \times 10^{-68}$			

Table 1: Test of immunosignaturing performance-This table lists values that were used to characterize immunosignature performance. From left to right: M represents the number of masks; Avg. CC Ab represents the average correlation coefficient calculated across every possible combination of the 51 monoclonal antibodies; Ave CC serum is the average correlation coefficient calculated across every possible combination of control and infected sample; ANOVA *p* -val is the *p* -value from the F-test using every monoclonal as a class, and the technical replicates as the measurements; Avg. raw value Ab is the average of all 51 monoclonals for each of the tested mask libraries; Median raw value is the median of all 51 monoclonals for each of the tested mask libraries; CV is the Coefficient of Variation (standard deviation corrected by mean) for every peptide across every monoclonal for each mask set. Note that the M=323 library was only used for the serum experiments

In order to examine the chemical characteristics of each library of 100,000 peptides, we used ProtParam (http://web.expasy.org/protparam/). A library of 100,000 peptides was deemed sufficient to summarize the trends for each mask set. ProtParam takes as input one-letter amino acid sequences and then predicts the chemical properties of each peptide. We examined molecular weight and theoretical pI to gauge how our generated peptides differed from completely random peptides, and we measured the number of repeated 2mer, 3mer, 4mer, 5mer and 6mer sequences to estimate library diversity. Each mask set is summarized by the average molecular characteristics of 100,000 peptides.

Peptide synthesis and array printing

The second part of the project is a biochemical test with actual peptides. Since 100,000 synthesized peptides would cost millions of dollars and many months to synthesize, we chose to use a much smaller library and increased the number of antibodies tested on these arrays to 51 very disparate monoclonals. Small-scale peptide synthesis yielded enough peptide to print thousands of microarrays. 384 peptides per mask set were synthesized at 80% purity, 5 mM scale by Sigma Genosys (St. Louis, MO). Failed syntheses were repeated until Sigma deemed the synthesis successful using their quality metrics. MALDI spectrograms indicated that each peptide had a major peak at the expected mass corresponding to an average purity of >92% according to Sigma Genosys. S1a and S1b Figures show an example of the MALDI chromatograms of a partial truncation that was resynthesized (S1a Figure) and a properly synthesized peptide (S1b Figure). Peptides were synthesized C-term to N-term, with a terminal Glycine-Serine-Cysteine linker that enabled attachment of the free N-term cysteine to a sulfo-SMCC linker (sulfosuccinimidyl 4-N-maleimidomethyl) cyclohexane-1-carboxylate, catalog number 22322, Pierce/Thermo Fisher Inc.) to a Schott A+ Nexterion aminosilane slide (catalog number 1064875, Schott, Jena, Germany). Truncated peptides were capped by the manufacturer and would not have a terminal cysteine ensuring that they do not attach to the slide. Peptide stocks were made by diluting 2 mg of peptide in distilled water + 20% acetonitrile at a final concentration of 2 mg/ml, based on measurement of peptide quantity by Agilent BioAnalyzer 2100 in protein mode (Agilent Inc., Santa Clara, CA). Microarray spotting plates were prepared using 384-deepwell plates by diluting the stock 2 mg/ml peptides into HEPES pH 7.3 with 2% TCEP (tris(2-carboxyethyl) phosphine) to a final concentration of 0.5 mg/ml. Printing was done at Applied Microarrays (Tempe, AZ) by piezo non-contact printing of approximately 150-200 pL per feature. Post-dispense, slides were stored at RT in 80% humidity to allow the maleimide conjugation reaction to occur. Prior to use, the arrays were washed with 30% ethanol, 500 mM EDTA pH 7.8, 40% acetonitrile to remove unbound peptides (Immunosignature effects paper). Every peptide was printed twice per assay. By duplicating the spots we ensured that any discrepancy in printing could be measured. The duplicate peptides were averaged for analysis. We used a 2-up assay system from Tecan (HS 4800 Hybridization Station, Tecan, Månnedorf, Switzerland) to perform the antibody incubations. The automation of the HS 4800 reduced operator and day-to-day variability. For every monoclonal, two technical replicates were run on different slides. Averaging across replicates accounts for a possible discrepancy in the assays. The technical replicates were run on physically different slides to reduce systematic bias in the assay. The technical replicates must exceed 0.85 correlation coefficient in order to achieve our quality threshold. Assays that failed would have been re-processed. However, during the experiment no replicate failed to meet our quality threshold. Post-assay measurements yielded peptide-to-peptide correlation average >0.99 suggesting that peptide printing was an insignificant source of variation. The average correlation across technical replicates was 0.96 suggesting that day-to-day and operator variance were insignificant sources of variation. Each microarray was incubated with one of the 100 different commercially sourced monoclonals listed in Table 1.

Immunosignature assays-monoclonals

For the monoclonal assay, only M=35, M=70, M=140 and M=272 peptide libraries were used. M=323 was only tested against human serum. Each of the 51 commercial monoclonals was added to 800 uL incubation buffer (3% BSA, 1x PBS pH 7.2, 0.05% Tween 20) to a final concentration of 5 nM. Incubation buffer + monoclonal was added to each microarray of 384 peptides. The Tecan HS4800 was programmed to incubate each slide with shaking for 1 hr at 37°C. The wash program included 3 washes of 5' each with 1x Tris-buffered saline pH 7.3 followed by 3 washes of 5' each using distilled water. When complete, the HS 4800 dried each slide with compressed nitrogen. Detection was by goat anti-Mouse IgG (H+L) directly conjugated to Alexafluor-555 (Life Technologies/Thermo Fisher) incubated at 4 nM for 1 hr followed by identical wash steps. Images were recorded by an Agilent 'C' scanner. Slides were scanned at 100% PMT, 70% laser power at 10 µm resolution using the 545 nm laser. Arrays were aligned using GenePix 6.0. Data was analyzed in R (CRAN.org) and GeneSpring 7.3.1 (Agilent, Santa Clara, CA).

Immunosignature assays–human sera

Eight patients who were diagnosed with Coccidiomycoses [11] with a measured titer of 1:256 against coccidioidin were tested against eight otherwise healthy age- and gender-matched volunteers in a standard immunosignature assay [11]. We measured the minimum *p*-value obtained between case and control for each mask set. All immunosignature assays were run in duplicate and averaged. T-tests were performed across all 384 peptides for each mask set and recorded. For M=323, a 10,000 random peptide microarray was used (NCBI GEO platform accession GPL17600), and 384 peptides were selected at random for analysis.

Analysis methods

In order to compare how effective different peptide cohorts work for immunosignatures, we first made the assumption that signatures for different monoclonal antibodies must be different [7]. We therefore calculated the Pearson's Correlation Coefficient for each set of 384 peptides across each of the 51 different monoclonal antibodies. If every monoclonal was mathematically distinguishable

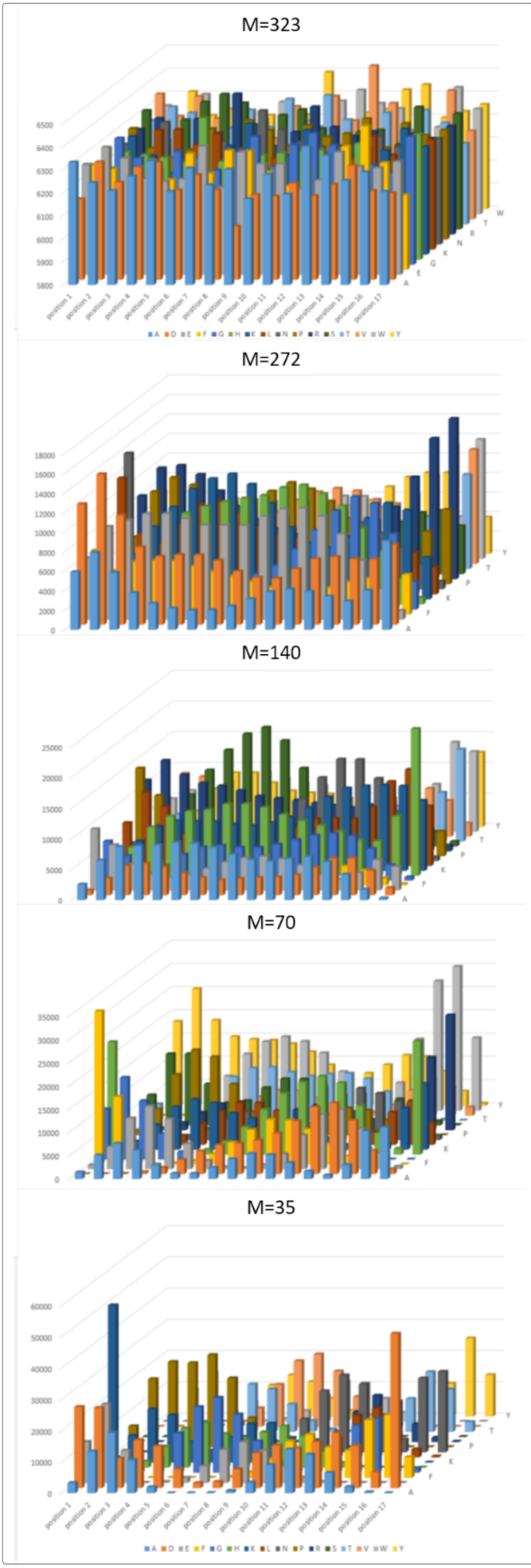

Figure 2: Left 5 panels: distribution of duplicated 5-mers for M=35, M=70, M=272 and M=323.

As the number of masks decreases, diversity in the resulting sequences is reduced, creating redundant sequences. The 5 panels above illustrate the number of duplicate 5mers. The X-axis is the total counts of duplicated 5mers in the 100,000 peptides, the Y-axis is the proportion of 5mers. The average for M323 (far right, top) is 1. The average number of duplicated 5mers for M=35 is 2 suggesting that redundancy is a hallmark of reduced mask numbers. Right 5 panels: Analysis of predicted peptide characteristics as a function of mask number. As in the left panels, each mask set created a distribution of data. Here the data is Molecular Weight. The trend seen here is similar for Extinction coefficient, beta sheet tendency, Alpha helix tendency, Beta turn tendency, Antigenicity, Isoelectric point, Aliphatic character, Average residue volume, Hydrophilicity, Flexibility, Sequence complexity, and Accessibility (all taken from ProtParam). X-axis is MW, Y-axis is the proportion of the total library for each MW value. The lower mask numbers tend to create sets of peptides with very specific and extreme molecular weights. For M35, weights ranged from 1856 to nearly 2600 while for M=323, weights ranged from 2380 to 2418. High mask numbers tend toward a more Gaussian distribution of molecular weights.

using immunosignatures, it follows that the correlation coefficients across the different antibodies would be low. Conversely, if the monoclonals could not be distinguished, or the information content per peptide library is low, it is likely that the correlation coefficients would be high. We did this for the 4 different mask sets from M35 to M272. This data is listed in Table 1.

For the coccidiomycoses set, we assayed eight different case and eight different control serum samples according to previous protocols [11]. We report the minimum *p*-value for each mask set, M=35, 70, 140, 272 and 323. For M=323, we performed the immunosignatures on the CIM 10,000 random peptide microarrays, and selected 384 peptides at random for the analysis. This prevented having to re-synthesize completely random (M=323) peptides, since completely random peptides already exist on the 10,000 peptide printed microarray. All peptide lists and data are publicly available at GEO (series accession numbers GSE50044, GSE50045).

In order to assess how peptide disorder affected immunosignature performance, we calculated Shannon's Entropy for each of the peptide libraries, since this method is a direct measure of disorder [44]. We used Support Vector Machines as the classifier for predictor of the identity of each monoclonal using a train/test paradigm. We mathematically removed 38 peptides from the 384 peptide libraries, and trained the SVM classifier. We then gave the classifier the left-out 38 peptides and asked it to predict the identity of each of the 51 different monoclonals. Thus, the classifier is completely naïve to the predictor peptides. This provided a way to directly assess peptide entropy *vs.* immunosignature performance. We used Shannon's Entropy formula:

$$y = \sum_{p=x,1-x} -p * log2(p)$$

Entropy for each peptide library was calculated per peptide and averaged.

Results

Informatic analysis of 100,000 virtual peptides for M=35, 70, 140, 272 and 323. We first wished to test the general characteristics of the 100,000 virtual peptides created by the Peptide Generator. Figure 1 shows the distribution of each amino acid at each position along the virtual peptides for each of five different peptide libraries. For M=323, each position was selected by a random number generator, not by Peptide Generator. For M=323 only 16 amino acids are shown, in order to match the other charts in Figure 1. The position along the peptide is shown on the X-axis and the amino acid prevalence at that position is shown on the Y-axis. The 16 amino acids are plotted along the Z-axis, alphabetically from front to rear. Colors are used to help readers follow the pattern for a given amino acid. Predictably, the M=323 set shows the least bias. M=323 is not perfectly even, however. The relative differences in prevalence for M=323 can be accounted for by random chance over just 100,000 peptides. Had more peptides been used, the differences from position to position would appear less. M=35 shows extreme biases as expected, but even M=272 shows a moderate cyclic bias caused by the nature of Peptide Generator's selection algorithm.

We then measured redundancy in the peptides that compose each library. Figure 2a, far left five panels, shows distribution plots indicating the number of duplicated 5mers for each of the five different libraries. M=323 has zero duplicated 5mers per peptide, while M=35 has on average two duplicated 5mers per peptide, indicating the relatively high redundancy imposed by the severe restriction in masks/synthesis steps. In order to calculate physical parameters that play a role in biochemical reactions, every peptide was processed by exPASy's ProtParam (www.expasy.org/protparam). The results from ProtParam were compiled and plotted (see Figure 2b, right five panels). The distribution of molecular weights for each library is shown for each mask set. As seen, the M=323 mask set yields an average molecular weight of 2391, with a standard deviation of 32. However, even taking into account that only 16 amino acids were used, rather than 19, the weight distribution for every mask restriction set shows a pronounced impact on the distribution of molecular weight. The broader the peak, the more times identical sub-sequences are used, contributing to more extreme molecular weights. The most often repeated 2mer, 3mer, and 4mer sequences are present, the more likely a peptide will be unusually heavy or light. This near-identical pattern was also seen for pI and other chemical peptide characteristics as calculated by ProtParam.

Figure 3 shows a more explicit analysis of redundancy, including 2mer, 3mer, 4mer, 5mer and 6mers. The more random a peptide library, the more unique sequences are present. The less diverse a library, the more repetition. The fact that the M=272 curve is just below M=323 suggests that M=272 may have only slightly more redundancy than M=323.

Biochemical analysis of actual peptides

In order to gain more perspective into these virtual libraries, we analyzed only the 16aa libraries (not M=323) in the following figures.

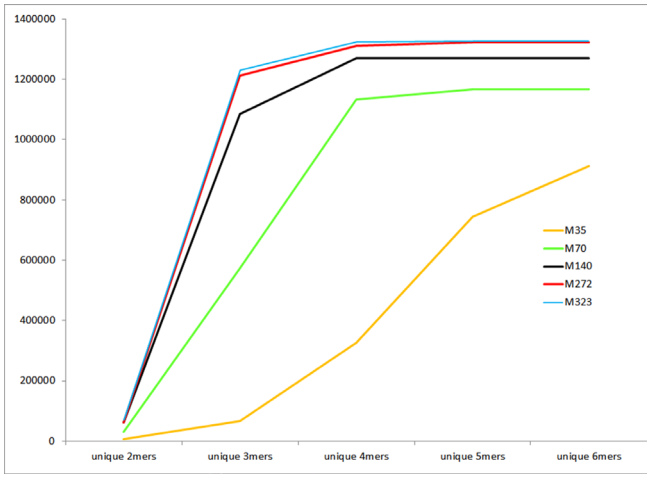

Figure 3: Display of unique 2mer, 3mer, 4mer, 5mer and 6mers contained within 100,000 peptides using the M=35, M=70, M=140, M=272 and M=323 mask restriction sets.

The Y-axis is the proportion of unique nmers, and the X-axis is increasing length of subsequences.

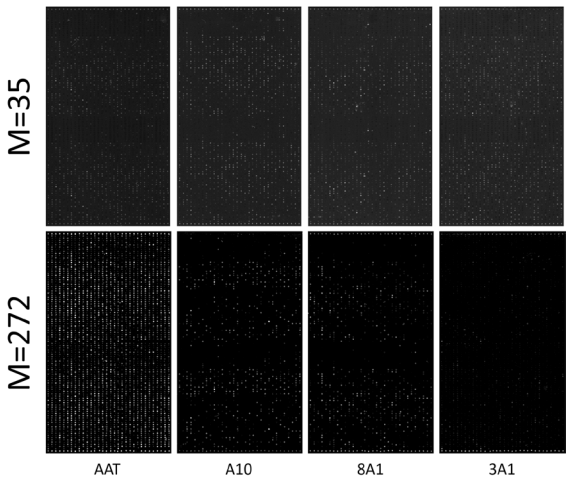

Figure 4: From left to right Top: Four images from antibodies AAT, A10, 8A1 and 3A1, the same as those used in subsequent figures.

Bottom: Same monoclonals on the M=272 library. These images show a promiscuous monoclonal that binds many peptides (AAT), two different moderate-specificity antibodies (A10 and 8A1) and a high specificity monoclonal that binds very few peptides but with high intensity (3A1). These four antibodies produce very different binding profiles on M=272 while M=35 produced very similar binding profiles.

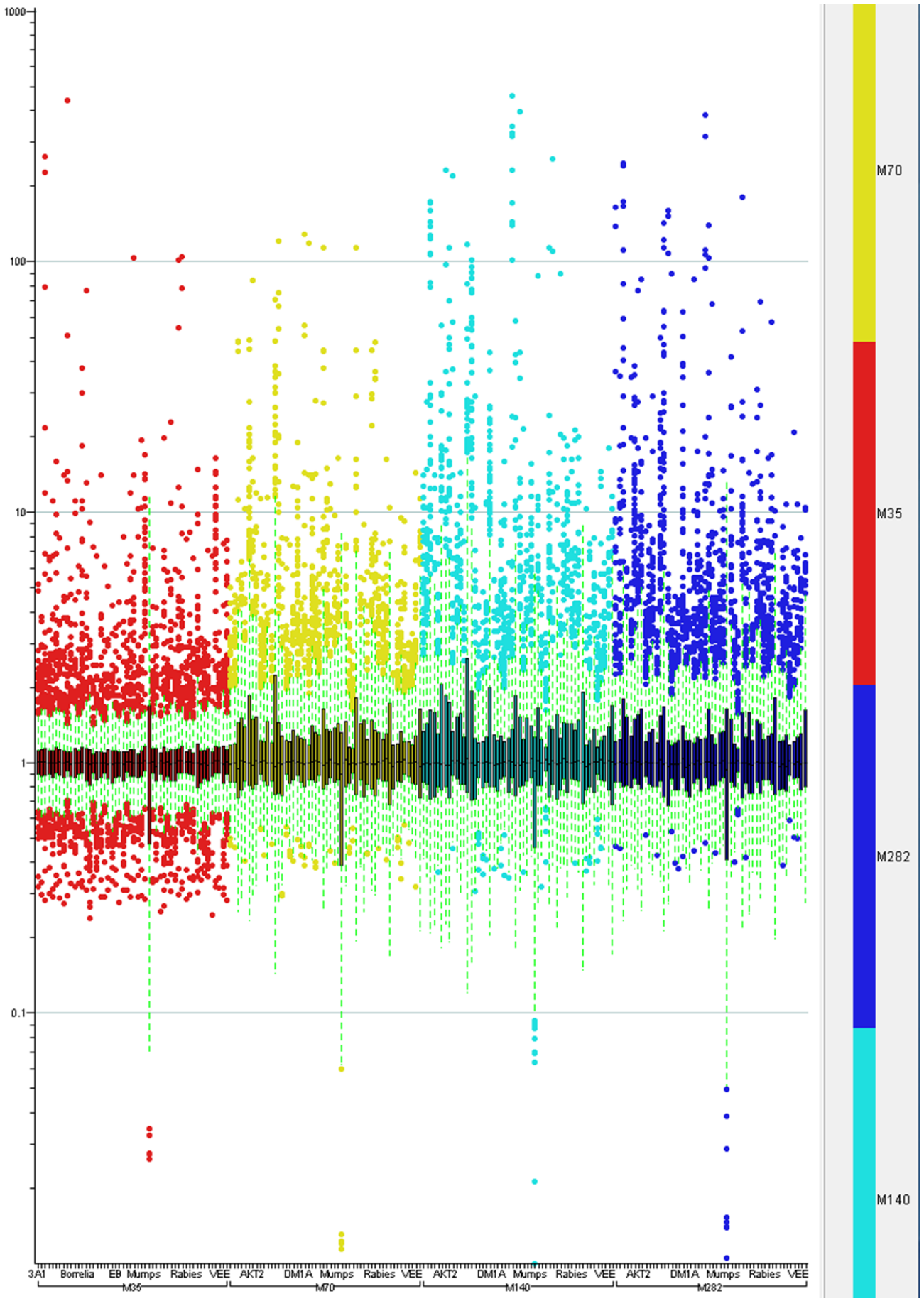

Figure 5: Bar graphs of each mask set (M=35, red; M=70, yellow; M=140, cyan; M=272, blue) for all of the 51 tested monoclonals.

Each dataset was median normalized, ensuring that the median signal is 1 for every monoclonal. The solid box is the first quartile, centered on the median. The dashed lines are the 95% percentile, outliers are show as colored dots.

Images of the actual arrays can be seen Figure 4, which compares two different libraries, M=35 and M=272, using monoclonal 3A1 (anti recombinant human Thioredoxin from Abcam), 8A1 (anti-human Thioredoxin, full-length) from Abcam, A10 (anti-human Thyroid-Stimulating Hormone Receptor from Abcam) and AAT (anti-Alpha-1-antitrypsin precurson, SERPINA1 from Santa Cruz Biotechnologies). Even by eye, one can see that the M=35 library has fewer unique signals arising from binding between antibody and peptide, while M=272 has a very diverse profile that varies substantially depending on the monoclonal. These images convey, at least on a gross level, that M=35 cannot distinguish between a low-specificity antibody like AAT (AAT binds many different peptides at equal intensities on a purely random peptide library) from a more specific antibody like 3A1 (which binds very few peptides on a purely random peptide library). These microarray images are converted into numerical data that is analyzed further.

Figure 5 is a view of the distribution of the median-normalized values for each of the 51 monoclonals across each of the four mask sets. Median normalization brings the median of each library's data distribution to 1. As seen here, M=35 has a greatly reduced interquartile range (solid box) and 95%ile range (dashed lines) *vs.* M=70, M=140 and M=272. M=272 appears very similar to M=140 in this presentation, leading one to believe that there is little to distinguish M=272 from M=240.

Figure 6 is a heatmap that examines four of the 51 total monoclonals across four of the mask libraries: 3A1, 8A1, A10 and AAT were all processed on M=35, M=70, M=140, and M=272. Each peptide was converted from a sequence to a number, from 1 to 384, and then sorted from lowest to highest raw intensity so that every peptide library can be plotted on the same heatmap. The rows in Figure 6 represent the non-parametric rank-order of the raw intensity value for each of the 384 peptides for the four monoclonals. This plotting method allows the intensity distributions to be more easily seen. The vertical colored bars (red, yellow, cyan, blue, and purple) represent the *k*-means (*k*=5) clustering of the values, with the purple group containing the highest intensities and the red group containing the lowest intensities. The heatmap illustrates two important characteristics of the monoclonals and the peptide libraries: first, the *total number* of peptides that each of these four monoclonals bound is different. This highlights the diversity in how specific or promiscuous each antibody can be relative to the peptide library. Second, the *range* of binding intensities decreases substantially as the restriction on the masks is tightened. M=35 shows very little diversity compared to M=272. This illustration expands on information seen in Figure 5.

Figure 7 is a principal components map (PCA) showing the first two variance components for all of the peptide libraries and all monoclonals (dots on the PCA plot). Variance component 1 is plotted on the X-axis and variance component 2 is plotted on the Y-axis. The PCA plot displays relative differences in total variance of the dataset along each of the axes. Red corresponds to M=35 and is centered on the map with generally low variance from monoclonal to monoclonal. Blue is M=272 and is spread out evenly across the both of the two principal components, while M=140 is spread out across only component 1 (X-axis). This further suggests that M=272 has an advantage in information content and specificity *vs.* M=35, M=70 and M=140.

Figure 8 directly compares the disorder or randomness inherent in each peptide library with the predictive capacity of that library to correctly predict each of the monoclonals. First, Shannon's Entropy was calculated for each peptide library. This value is plotted on the Y axis. Next 38 peptides were removed from each 384 peptide library. 346 peptides were used to train a Support Vector Machine classifier, and the 38 removed peptides were used to test the SVM classifier. Since the classifier is predicting the identity of the monoclonals using untrained, it is unlikely to predict perfectly. This was done 10 times, using a different set of removed peptides each time. The accuracy of the monoclonal predictions was averaged and is plotted on the X-axis in Figure 8. This figure directly compares the peptide randomness against immunosignature performance. R2 is 0.7029 suggesting that immunosignature performance drops off more quickly than peptide redundancy increases (Table 1).

Discussion

We provided a rationale for using random peptides for immunosignatures, and a means for optimizing *in situ* manufacturing of random peptide microarrays. We examined a computational and a biochemical method for comparing how random and pseudorandom peptides differ. We created 100,000 virtual peptides using an algorithm that creates peptides when M (synthesis steps) is <AA (# of amino acids) * Length (of peptide). When M<AA*L, peptide sequences are no longer random. They acquire a bias caused by the lack of freedom in choosing any amino acid during synthesis. The fewer the synthesis steps (or masks, when using shadow-mask lithography), the more pronounced the bias. We first noted an impact on the chemical characteristics for peptides produced when M=35, M=70, and M=140. When M=272 and M=323, the redundant peptide sequences were substantially less than the other libraries. Redundancy in M=35 was extremely high. Average molecular weight and pI had a very high standard deviation in the low-M libraries because redundant amino acids and repeated 3mers or 4mers caused peptides to have many extremely high or extremely low weights and pI values. When amino acids are chosen at random, stretches of similar 3mer or 4mers are

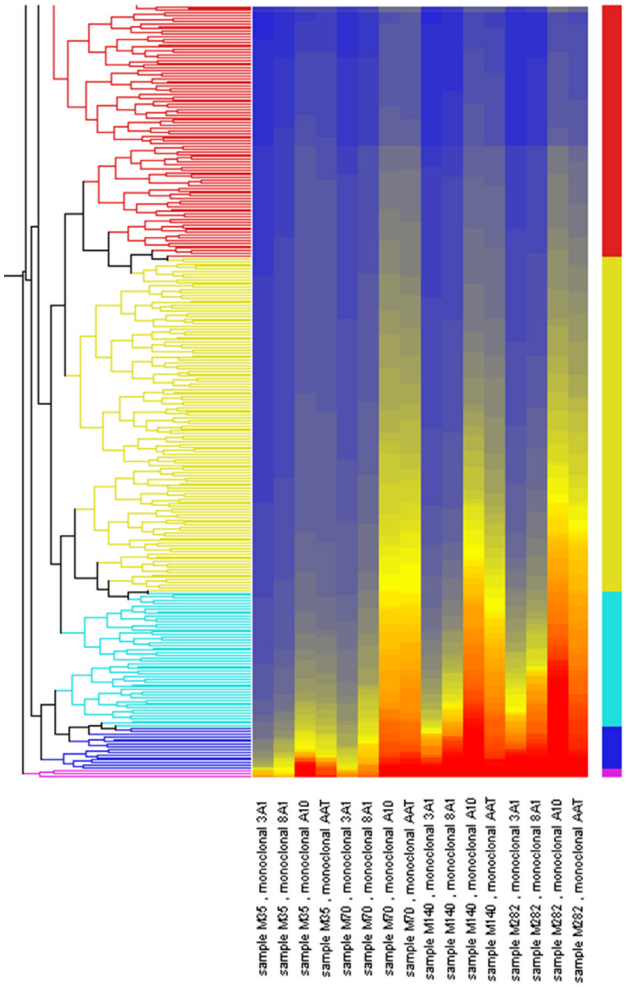

Figure 6: Heat map four of the 51 total monoclonals across four of the mask libraries: 3A1, 8A1, A10 and AAT, all processed on M=35, M=70, M=140, and M=272.

Each peptide was converted from a sequence to a number, from 1 to 384, and sorted from lowest to highest. The rows represent the rank-order of the intensity values for each of these four monoclonals. The vertical colored bars (red, yellow, cyan, blue, and purple) represent the k-means clustering of the values, with the purple group containing the highest intensities and the red group containing the lowest intensities.

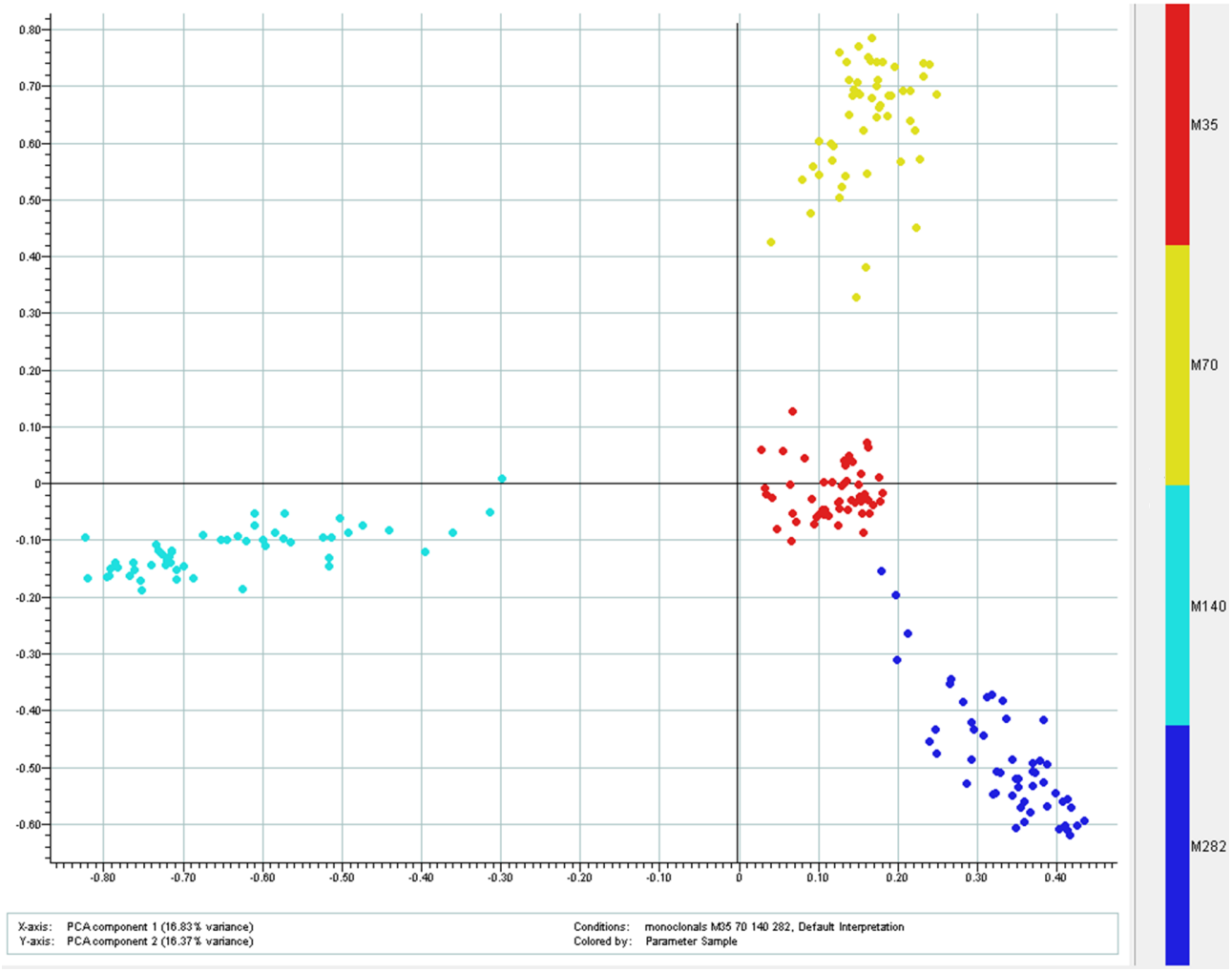

Figure 7: Principal Components Map (PCA) of all 51 monoclonals (colored dots) for each of the four mask sets (colored by mask set). The first two principal components are plotted on the X and Y-axes, respectively. The least dispersed is M=35, the most M=272 (as it is spread across both component 1 and component 2).

nearly undetectable, but as mask numbers decreased these redundant 3mer and 4mer sequences skewed the weight and pI distribution. Even 5mer redundancy was pronounced in M=35 and M=70, less so in M=140 and M=272. Without the full freedom to select any amino acid at any position, sequence redundancy actually becomes readily at M=140 and increases nearly logarithmically by the time M=35.

These results did not indicate whether an immunosignature would work, since previous publications used completely random peptide sequences, they merely showed the extent to which physical properties change when amino acids are constrained. It was altogether possible that peptides with extreme molecular weights and pI values would work as well or better than completely random peptides. We therefore tested a set of monoclonal antibodies and human serum on peptide libraries using different mask restrictions. We synthesized a set of peptides using M=35, M=70, M=140 and M=272 mask restrictions. A completely random peptide library already existed in the form of a 10,000 peptide microarray, where M=323 (GEO accession 17600), so a set of 384 peptides was randomly selected as M=323. This set was used to test human serum, but we could not supply enough 10,000 peptide arrays to complete every monoclonal experiment. The trend was clear: each mask restriction showed decreasing performance relative to distinguishing the 51 different monoclonals. Previous reports indicated that the 10,000 random peptide microarrays could easily distinguish monoclonals using less than 100 ANOVA-selected peptides [7]. This was the high-bar against which the other peptide libraries would be compared.

Figure 4 suggests at a gross level that M=35 is far too constrained to enable adequate immunosignature performance. The signals for different monoclonals become qualitatively monotonic. The box plots in Figure 5 suggest that the dynamic range of the restricted peptide libraries are greatly affected, as seen in the upper and lower quartile boxes and the 95th percentile lines. This does not yet imply the performance is worse, but clearly shows a drastic change in the data distributions. More telling is the heatmap in Figure 6. The range of values is extremely restricted for each monoclonal particularly for M=35. This low dynamic range suggests that values for each monoclonal are limited to very few high intensity values, but mostly low values. This may be insufficient for informative immunosignatures. The PCA plot in Figure 7 illuminates an interesting phenomenon: M=35 and M=70 are both clustered tightly near the center of the plot, which would be predicted as these are the most constrained. Compare M=272 which is dispersed across the first two principal components and M=140, which is dispersed only across the first principal component. It might be more likely that the variance in M=140 comes from less diverse patterns of reactivity than M=272. M=35 and M=70 are so invariant that they should be removed from consideration as useful sets.

M=140 demonstrated other interesting behavior. Figure 3 shows very little apparent difference between M=140 and M=272 in terms of internal sequence redundancy, on a gross level. The serum comparison (Table 1) suggests M=140 is similar in performance to M=272. M=140 might be considered as the reduction in shadow masks drops upfront costs from $1,600,000 to $200,000, and synthesis time from 2 weeks to 4 days.

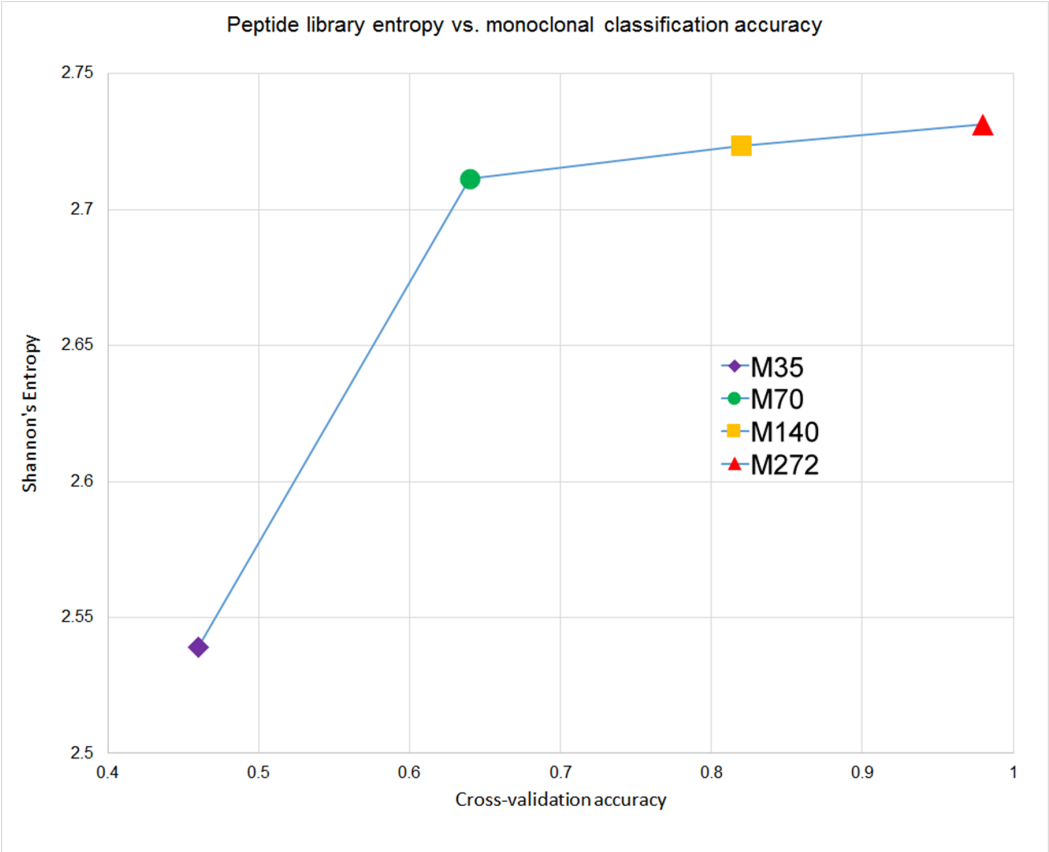

Figure 8: Shannon's Entropy for each of four peptide libraries *vs.* cross-validation accuracy of prediction of monoclonals Shannon's Entropy, a standard measure of disorder, was calculated for each of four different peptide libraries (Y axis).

Each of 50 different monoclonals was predicted by Support Vector Machines using a randomly selected 1/10 of the total 384 peptides per library. This was done 10 times, and the average accuracy for each peptide library is plotted on the X axis. R^2 for a linear fit is 0.7029.

To this point, we have not directly compared immunosignature performance with peptide randomness. Figure 8 shows a comparison of the disorder in the peptide libraries as measured by Shannon's Entropy *vs.* the accuracy in predicting the identity of the 51 monoclonals using Support Vector Machine classifier. Entropy is a measure of disorder. Any predictable pattern in a given dataset means the data is non-random, and entropy would be less than the maximum. Here we see the rapid change in entropy between M=35 and the next library, M=70, with much smaller jumps between the subsequent peptide sets. The R^2 of the linear regression is 0.7, which suggests that the redundancy in the peptide libraries increases non-linearly as the number of synthesis steps drops by nearly 1 log. Since M=35 is unlikely to produce a useful peptide library, given the substantial sequence redundancy, we will focus instead on M=70, 140 and 272. The R^2 of the regression line for these three libraries is 0.9936, suggesting that randomness is still measurably different across these libraries. Shannon's Entropy for M=323 is 2.93 implying that there is less than maximum disorder in the M=272 library. This is likely due to the Peptide Generator's greedy algorithm, which is not explicitly tasked with selecting amino acids at random, but rather probabilistically per synthesis cycle.

In order to estimate immunosignature performance, we removed 1/10 of the peptides in each library, and held that data back. We then trained SVM with the remaining peptides and asked it to predict every monoclonal based solely on the pattern of binding to the 38 held-back peptides. If the dataset were monotonic, prediction accuracy would suffer. If every peptide were informative and unique, then the held-back peptides would contain information about the monoclonals that is unrelated to the training peptides, and could therefore independently predict the identity of the monoclonals. The small increase in entropy between M=140 and M=272 resulted in a large boost in prediction accuracy, from 82% correct to 98% correct. This may be the most critical measure in mask restriction, since monoclonals are much easier to classify than human patients with a disease and that boost in predictive power could be even more important when measuring billions of antibodies in serum.

Conclusions

The data provided here can be applied to maskless peptide syntheses as well, by reducing the number of steps. In cases where amino acids are added simultaneously to growing chains of peptides in multiplexed reaction vessels, the number of cycles can be reduced. As the number of peptides simultaneously synthesized increases, the value of restricting the number of steps increases. Perhaps most importantly, it is not necessary to use the output of Peptide Generator as-is. A more useful method is to produce every possible peptide for a given mask restriction. Then, filter this large collection of peptides. The user can decide what filter is most important, whether reducing redundancy, reducing certain physical characteristics such as hydrophobicity, or picking random peptides with the most life-space motifs. Many of the biases described in this manuscript can be moderated by this post-hoc filter. Depending on the assay, performance of a restriction-set of peptides, if properly filtered, might be competitive with a fully random set. The assay drives the balance between reducing synthesis steps and retaining peptide diversity and disorder. One of the quickest methods for improving the utility of a random peptide library is reducing 3mer, 4mer and 5mer redundancy. Even without restriction, a random library groomed for minimal 5mer repeats is more likely to contain more unique linear targets than an unfiltered set. That single change could enhance epitope identification by an order of magnitude or more. Judicious selection of a library of mask-restricted peptides would not only

improve the performance of the resulting random peptide library, but the impact to synthesis time and cost would be substantial.

Although this manuscript focused on shadow-mask lithography, our results can be applied to many peptide synthesis methods. Random peptide libraries are quite useful for a number of immunological and biochemical techniques. As peptide microarrays become denser, the advantages of using random microarrays over phage display grows. Although there are many ways of making peptide microarrays, costs are still very high and production volume remains relatively low. In order to use these arrays for any sort of diagnosis, the costs must drop orders of magnitude and production volume must scale up as well. Shadow mask synthesis can facilitate extremely high throughput production, low cost synthesis equipment, and scalable costs that trend the same way semiconductor manufacturing does. This enables commodity-level high quality peptide microarrays, facilitating new applications in medical care, diagnostics, and even therapeutic antibody quality control and assessment.

Data are provided at the Gene Expression Omnibus at NCBI. Platform: GPL17490, GPL17679 and series Series GSE49217, GSE50044 and GSE50045, "*In situ* synthesis of peptide microarrays - shadow mask design".

Acknowledgments

There were no funding agencies for this work. Data are freely available at the Gene Expression Omnibus at NCBI. Platforms: GPL17600, GPL17490, GPL17679, Experimental Series: GSE49217, GSE50044, GSE50045.

References

1. Buus S, Rockberg J, Forsström B, Nilsson P, Uhlen M, et al. High-resolution Mapping of Linear Antibody Epitopes Using Ultrahigh-density Peptide Microarrays. Mol Cell Proteomics. 2012 Dec;11(12):1790-800.
2. Domenyuk V, Loskutov A, Johnston SA, Diehnelt CW. A Technology for Developing Synbodies with Antibacterial Activity. PLoS ONE. 2013 Jan;8:e54162.
3. Boltz KW, Gonzalez-Moa MJ, Stafford P, Johnston SA, Svarovsky SA. Peptide microarrays for carbohydrate recognition. Analyst. 2009 Apr;134(4):650-652.
4. Sinzinger MD, Chung YD, Adjobo-Hermans MJW, Brock R. A microarray-based approach to evaluate the functional significance of protein-binding motifs. Analytical Anal Bioanal Chem. 2016 Feb;408:3177–3184.
5. Houseman BT, Huh JH, Kron SJ, Mrksich M. Peptide chips for the quantitative evaluation of protein kinase activity. Nat Biotech. 2002 Mar;20:270-274.
6. Fu J. Microarray Selection of Cooperative Peptides for Modulating Enzyme Activities. Microarrays. 2017 Jun;6(2):8.
7. Stafford P, Halperin R, Legutki JB, Magee DM, Galgiani J, et al. Physical characterization of the 'Immunosignaturing Effect'. Mol Cell Proteomics. 2012 Apr;11(4):M111.011593.
8. Halperin RF, Stafford P, Legutki JB, Johnston SA. Exploring antibody recognition of sequence space through random-sequence peptide microarrays. Mol Cell Proteomics. 2011 Mar;10(3):M110.000786.
9. Geysen HM, Rodda SJ, Mason TJ, Tribbick G, Schoofs. Strategies for epitope analysis using peptide synthesis. J Immunol Methods. 1987 Sep;102(2):259-274.
10. Huang J, Ru B, Zhu P, Nie F, Yang J, et al. MimoDB 2.0: a mimotope database and beyond. Nucleic Acids Res. 2012 Jan;40:D271-D277.
11. Navalkar K, Magee DM, Galgiani J, Cichacz Z, Johnston SA, et al. Application of immunosignatures to diagnosis of Valley Fever. Clin Vaccine Immunol. 2014 Aug;21(8):1169-1177.
12. Restrepo L, Stafford P, Magee DM, Johnston SA. Application of immunosignatures to the assessment of Alzheimer's disease. Ann Neurol. 2011 Aug;70(2):286-295.
13. Singh S, Stafford P, Schlauch KA, Tillett RR, Gollery M, et al. Humoral Immunity Profiling of Subjects with Myalgic Encephalomyelitis Using a Random Peptide Microarray Differentiates Cases from Controls with High Specificity and Sensitivity. Mol Neurobiol. 2018 Jan;55(1):633-641.
14. Stafford P, Cichacz Z, Woodbury NW, Johnston SA. Immunosignature system for diagnosis of cancer. PNAS. 2014 Jul;111(30):E3072-E3080.
15. Rodi DJ, Soares AS, Makowski L. Quantitative Assessment of Peptide Sequence Diversity in M13 Combinatorial Peptide Phage Display Libraries. J Mol Biol. 2002 Oct;322(5):1039-1052.
16. Legutki JB, Zhao ZG, Greving M, Woodbury N, Johnston SA, et al. Scalable high-density peptide arrays for comprehensive health monitoring. Nat Commun. 2014 Sep;5:4785.
17. Richer J, Johnston SA, Stafford P. Epitope Identification from Fixed-complexity Random-sequence Peptide Microarrays. Mol Cell Proteomics. 2015 Jan;14(1):136-147.
18. Halperin R, Stafford P, Johnston SA. GuiTope: an application for mapping random-sequence peptides to protein sequences. BMC Bioinformatics. 2012;13:1.
19. Donnell B, Maurer A, Papandreou-Suppappola A, Stafford P. Time-Frequency Analysis of Peptide Microarray Data: Application to Brain Cancer Immunosignatures. Cancer Inform. 2015 Jun;14(Suppl 2):219-233.
20. Zwick MB, Shen J, Scott JK. Phage-displayed peptide libraries. Curr Opin Biotechnol. 1998 Aug;9(4):427-436.
21. Restrepo L, Stafford P, Johnston SA. Feasibility of an early Alzheimer's disease immunosignature diagnostic test. J Neuroimmunol. 2013 Jan;254(1-2):154-160.
22. Williams S, Stafford P, Hoffman S. Diagnosis and early detection of CNS-SLE in MRL/lpr mice using peptide microarrays. BMC Immunol. 2014 Jun;15:23.
23. Hughes A, Cichacz Z, Scheck AC, Coons SW, Johnston SA, et al. Immunosignaturing can detect products from molecular markers in brain cancer. PLoS ONE. 2012;7(7):e40201.
24. Stafford P, Wrapp D, Johnston SA. General Assessment of Humoral Activity in Healthy Humans. Mol Cell Proteomics. 2016 May;15(5):1610-1621.
25. Wang L, Whittemore K, Johnston SA, Stafford P. Entropy is a Simple Measure of the Antibody Profile and is an Indicator of Health Status: A Proof of Concept. Scientific Reports. 2017 Dec;7:18060.
26. Young JD, Huang AS, Ariel N, Bruins JB, Ng D, et al. Coupling Efficiencies of Amino Acids in the Solid Phase Synthesis of Peptides. Pept Res. 1990 Aug;3(4):194-200.
27. Gao X, Pellois JP, Na Y, Kim Y, Gulari E, et al. High density peptide microarrays. *In situ* synthesis and applications. Mol Divers. 2004;8(3):177-187.
28. Gao X, Gulari E, Zhou X. *In situ* synthesis of oligonucleotide microarrays. Biopolymers. 2004 Apr;73:579-596.
29. Fodor S, Read J, Pirrung M, Stryer L, Lu A, et al. Light-directed, spatially addressable parallel chemical synthesis. Science. 1991 Feb;251(4995):767-773.
30. Choung RS, Marietta EV, Van Dyke CT, Brantner TL, Rajasekaran J, et al. Determination of B-Cell Epitopes in Patients with Celiac Disease: Peptide Microarrays. PLoS One. 2016 Jan;11(1):e0147777.
31. Pellois JP, Zhou X, Srivannavit O, Zhou T, Gulari E, et al. Individually addressable parallel peptide synthesis on microchips. Nat Biotechnol. 2002 Sep;20(9):922-926.
32. Beyer M, Nesterov A, Block I, König K, Felgenhauer T, et al. Combinatorial synthesis of peptide arrays onto a microchip. Science. 2007 Dec;318(5858):1888.
33. Maurer K, McShea A, Strathmann M, Dill K. The removal of the t-BOC group by electrochemically generated acid and use of an addressable electrode array for peptide synthesis. J Comb Chem. 2005 Sep-Oct;7(5):637-640.

34. Agarwal PB, Pawar S, Reddy SM, Mishra P, Agarwal A. Reusable silicon shadow mask with sub-5 μm gap for low cost patterning. Sensors and Actuators A: Physical. 2016 May;242:67-72.

35. Kay BK, Adey NB, Yun-Sheng H, Manfredi JP, Mataragnon AH, et al. An M13 phage library displaying random 38-amino-acid peptides as a source of novel sequences with affinity to selected targets. Gene. 1993 Jun;128(1):59-65.

36. Noren KA, Noren CJ. Construction of High-Complexity Combinatorial Phage Display Peptide Libraries. Methods. 2001;23(20:169-178.

37. Cai C, Dai X, Zhu Y, Lian M, Xiao F, et al. A specific RAGE-binding peptide biopanning from phage display random peptide library that ameliorates symptoms in amyloid β peptide-mediated neuronal disorder. Appl Microbiol Biotechnol. 2016 Jan;100(2):825-835.

38. Ge M, Yan A, Luo W, Hu YF, Li R-C, et al. Epitope screening of the PCV2 Cap protein by use of a random peptide-displayed library and polyclonal antibody. Virus Res. 2013 Oct;177(1):103-107.

39. Chin CF, Tan SJ, Gan CY, Lim TS. Identification of Peptide Based Inhibitors for α-Amylase by Phage Display. Int J Pept Res Ther. 2015 Feb;21(3):237-242.

40. Kukreja M, Johnston SA, Stafford P. Immunosignaturing microarrays distinguish antibody profiles of related pancreatic diseases. J Proteomics Bioinform. 2012 Jan;S6:001.

41. Legutki JB, Johnston SA (2013) Immunosignatures can predict vaccine efficacy. PNAS. 2013 Nov;110(46):18614-18619.

42. Boltz KW, Nagaraj V, Svarovsky S, Lake D, Stafford P, et al. High throughput technology for the identification and characterization of glycan binding peptides. Glycobiology. 2006;16(3):1155-1156.

43. Kuznetsov IB. Identification of non-random sequence properties in groups of signature peptides obtained in random sequence peptide microarray experiments. Biopolymers. 2016 May;106(3):318-29.

44. Giancarlo R, Scaturro D, Utro F. Textual data compression in computational biology: a synopsis. Bioinformatics. 2009;25(13):1575-1586.