MCP Papers in Press. Published on November 9, 2010 as Manuscript M110.000786

Exploring Antibody Recognition of Sequence Space through Random-Sequence Peptide Microarrays

Rebecca F. Halperin, Phillip Stafford, Stephen Albert Johnston*

Center for Innovations in Medicine
Biodesign Institute
Arizona State University
PO Box 875901
Tempe AZ 85281

*corresponding author
Phone: 480-727-0792
Fax: 480-727-0782
Email: stephen.johnston@asu.edu

Running Title: Antibody Recognition of Random-Sequence Peptides

Abbreviations:

ROC: Receiver Operator Characteristic

KLH: Keyhole limpet hemocyanin

Summary

A universal platform for efficiently mapping antibody epitopes would be of great use for many applications, ranging from antibody therapeutic development to vaccine design. Here we tested the feasibility of using a random peptide microarray to map antibody epitopes. Although peptide microarrays are physically constrained to ~10^4 peptides per array, compared to 10^8 permitted in library panning approaches such as phage display, they enable a much more high though-put and direct measure of binding. Long (20mer) random sequence peptides were chosen for this study to look at an unbiased sampling of sequence space. This sampling of sequence space is sparse, as an exact epitope sequence is unlikely to appear. Commercial monoclonal antibodies with known linear epitopes or polyclonal antibodies raised against engineered 20mer peptides were used to evaluate this array as an epitope mapping platform. Remarkably, peptides with the most sequence similarity to known epitopes were only slightly more likely to be recognized by the antibody than other random peptides. We explored the ability of two methods singly and in combination to predict the actual epitope from the random sequence peptides bound. Though the epitopes were not directly evident, subtle motifs were found among the top binding peptides for each antibody. These motifs did have some predictive ability in searching for the known epitopes among a set of decoy sequences. The second approach using a windowing alignment strategy, was able to score known epitopes of monoclonal antibodies well within the test dataset, but did not perform as well on polyclonals. Random peptide microarrays of even limited diversity may serve as a useful tool to prioritize candidates for epitope mapping or antigen identification.

Introduction. Antibodies play an important role in protecting against infectious disease and contribute to pathology in autoimmune disease. Understanding antibody-antigen interactions is important for elucidating disease etiology, as well as facilitating vaccine design and diagnostic test development. In addition to their role in the immune system, antibodies are also very useful as affinity reagents for detection and purification as well as clinical diagnostic tools and pharmaceuticals. Epitope mapping is often an important step in determining if an antibody is suitable for a particular application, sorting among antibodies or determining how it performs its function. Many methods exist for identifying the epitope of an antibody, including crystallography, peptide tiling, and phage display (1-2). In this study, we will examine whether a faster, less expensive array based approach could be applied to the epitope mapping problem

A crystal structure of the antibody bound to the target is generally considered the gold standard of epitope mapping because it gives the most detailed information about the binding mechanism and will work for both conformational and linear epitopes. To identify a linear epitope, the peptide tiling method is often preferred because it is simple and straightforward. However, the expense of synthesizing tiling peptides for every protein target may be prohibitive. To avoid the costly synthesis step, a library approach such as phage display may be employed. Peptides with random sequences can be displayed on the surface of phage, and those that bind best to the antibody can be selected and amplified. In the case of a linear epitope, the sequences recovered generally have sequences that very closely or exactly match the epitope sequence (3-8). However, several rounds of selection, as well as sequencing of many selected clones makes this process expensive and time consuming. Furthermore, phage display has an inherent bias in selecting peptides that facilitate growth which reduces the effective size of the library. A faster method that allows a more direct measure of binding would be ideal.

Peptide arrays provide an alternative for screening a library of peptides for binding activity. The challenge of the array based approach is that the size of the peptide library feasible is several orders of magnitude smaller than those typically used in phage display. We have developed a random-sequence peptide microarray and are exploring its usefulness in a number of applications. The peptide library consists of 10,340 random sequence peptides that have seventeen randomized positions and a three amino acid linker. The library represents a very sparse sampling of sequence space, as only five percent of all possible 5mer sequences are represented. Despite the small library size, the random-sequence peptide microarray was successfully used to identify protein and glycan binding peptides, and most pertinent to this study, to profile humoral immune responses (9-12). This technique known as immunosignaturing is a novel method to detect changes in antibody reactivity has been described by our group elsewhere (12 and under preparation). The peptides need only be mimotopes for immunosignaturing to serve its main purpose as a diagnostic platform. However, it is obvious to ask how the peptide sequences may relate to the antigen that raised the detected antibody response. It would be very useful if the peptide sequences identified in the immunosignaturing experiments could be used to identify the immunogenic epitopes in a pathogen or an autoantigen. Some preliminary use of random peptide arrays for epitope mapping was done by Reinke et al., but required subsequent rounds of mutational analysis of the peptides to hone in on the epitope sequence (13). However, it may be possible to use a more sophisticated bioinformatics approach to infer the epitope from these loosely related sequences.

In order to evaluate how well the random-sequence peptide microarray could work for predicting the antibody epitopes, ten antibodies with known epitopes were selected. Five well characterized monoclonal antibodies that recognize linear epitopes were selected. Two

recognize distinct P53 epitopes, which had been identified through peptide tiling and phage display experiments (14). The antibody against tubulin was epitope mapped using protease digestions, and the anti-IL2 has been co-crystallized with its target (15-16). The anti-HA monoclonal was raised against the peptide commonly used as an epitope tag. Five polyclonal antibodies raised against random peptides from the library were included to evaluate the more difficult task of predicting an epitope from a polyclonal immune response. Using anti-peptide sera allows us to have a polyclonal response, but still know where the epitopes must be. We chose peptides from the library to facilitate comparison of the relative binding levels of the cognate peptide and the array of mimotopes.

Motif finding algorithms are able to find subtle patterns in sets of unaligned sequences. These algorithms may be classified in two main categories: deterministic and optimizing. Deterministic algorithms will exhaustively search a sequence set for motifs fitting a well defined set of criteria. Some popular implementations of deterministic motif finding algorithms are TEIRESIAS or PRATT (17-18). The optimizing algorithms represent the motif probabilistically and try to maximize a scoring function. The optimization can be preformed stochastically such as using Gibbs motif sampling or by expectation maximization as implemented in MEME (19). An optimization approach seems most appropriate for this problem because we do not know what criteria the motif should fulfill. Since it is possible that gapped motifs may be useful here, the GLAM2 implementation of the Gibbs motif sampling algorithm will be used here because it allows for gaps (20).

An alternative to finding a motif among the peptides would be to compare the peptides one at a time to the antigen sequence(s). A similar sequence analysis problem was addressed by a group using peptides selected by phage display to bind to small molecules to identify

analogous binding sites in real proteins. The algorithm implemented in the RELIC MATCH program compares each peptide sequence to the target protein sequence in five amino acid windows, and scores each window for similarity (21). The scores for all of the peptides are added up across the protein sequence to predict potential small molecule binding site. A similar approach may be useful for predicting antibody recognition sites from dissimilar peptide sequences selected in a peptide microarray experiment.

Typically, epitope mapping is performed in order to identify the specific part of a protein target that is recognized by the antibody. A recent study has demonstrated the feasibility of using a similar approach to identify an unknown protein target of an antibody (22). Peptides from a phage display library can be selected against an antibody. Motifs in the peptide sequences can be used to search a database of potential antibody targets. The authors concluded that a motif of at least seven amino acids or two shorter motifs in combination could be used to reasonably identify a protein target among a database of candidates. This approach could be powerful in identifying the antigenic proteins in a pathogen, targets in an autoimmune disease, or even discovering the cause of an unknown infection and the potential to translate this approach to a microarray platform will be explored here. Here we determine to what extent a bioinformatics approach may be used to predict the epitope directly from the 10K peptide array data.

(Figure 1)

Methods. Random sequence peptide arrays were produced as described in Morales Betanzos et. al. (11). Briefly, 10K random 17mer sequences containing equal probability of the 20 amino acids except cysteine were generated. A C-terminal linker of Gly-Ser-Cys was added to each

sequence to make 20mer peptides. These peptides were synthesized by Alta Bioscience (Birmingham, UK), diluted in 25% DMF 75% PBS, and spotted on a Telechem Nanoprint quill type contact printer onto sulfo-SMCC activated aminosilane coated glass slides. *(Table 1)*

Antibodies with known epitopes were purchased from commercial suppliers. mAb1, mAb2, mAb3 and mAb4 were purchased from Labvision (Fremont, CA) and mAb5 was purchased from Abcam (Cambridge, MA). The five polyclonal antibodies were produced as follows. Mice were immunized with keyhole limpet hemocyanin (KLH) conjugated peptides and sera was obtained at day 35. All animal work was conducted following an animal use protocol which was approved by the Arizona State University Institutional Animal Care and Use Committee. Antigen specific antibodies were absorbed from sera by binding to KLH immobilized on nitrocellulose membrane. A one by six centimeter nitrocellulose membrane was placed in a 15 ml conical tube with 1.0 mg/ml KLH in 2.0 ml PBS. The membrane was washed three times in TBST and incubated with 1.0% BSA in TBST for at least one hour or until used. After washing three times in TBST, the membrane was placed into 2.0 ml of sera diluted 1:500 in 3% BSA, 0.05% Tween, PBS buffer. Reactivity of sera for KLH was tested in an ELISA. Sera were considered absorbed when no reactivity for KLH was detected at the 1:500 dilution.

The monoclonal antibodies or polyclonal sera were used to probe the peptide microarrays. After slides were passivated with 0.014% Mercaptohexanol, antibody was diluted to 100nM or sera were diluted 1:500 in 3% BSA, 0.05% Tween, PBS. Antibodies were incubated with slides for 1 hour at 37C in Agilent Chambers with rotation. Slides were washed three times with TBS, 0.05% Tween and three times with diH_2O. The incubation and wash procedure was repeated with a biotinalyted secondary antibody (Bethyl Laboratoreis, Inc.

Montgomery, TX), then with Alexa-555 labeled Streptavadin (Invitrogen, Carlsbad, CA). Negative control arrays with no primary antibody or naïve mouse sera were also run for comparison. At least three replicate arrays were run for each antibody.

Slides were aligned using GenePix Pro 6.0 (Molecular Devices) and median spot intensities were averaged across replicate slides. Negative control signals were subtracted from antibody signals to remove the contribution of the secondary binding. The top 500 peptides in fluorescent intensity were selected for each antibody. The number of times each peptide occurs in one of the top 500 peptides lists was tabulated. Peptides appearing in five or more lists were eliminated as they are likely Fc binders or other nonspecific interactions. For the polyclonal datasets, the immunizing peptides were also excluded. The peptide lists described here will be referred to as the *binders* lists in subsequent analysis.

Peptides from the array were compared to the epitope sequences to identify those with sequence similarity. The epitope was expressed as a GLAM2 motif and was used in GLAM2SCAN to search against the peptides from the array inserted in strings of cysteines, with an alphabet of equal amino acid frequencies. Peptides were sorted by the highest scoring match and lists of the best matching peptides were created and these lists are referred to as the *aligners* lists. These lists were compared with lists of peptides that most strongly bind to each peptide and the proportion of overlap was examined.

Test datasets were generated for the monoclonal antibodies by randomly selecting sequences from human Swissprot and then randomly selecting a window of that sequence the same length as the epitope sequence. Two hundred negative examples were generated for each monoclonal. One thousand random peptides were generated as the negative examples for the polyclonal antibodies with equal frequencies of the nineteen amino acids (cysteine was not

included as in the arrays). All of these sequences were inserted within a string of seventeen cysteines on each side to allow peptides to be aligned overhanging the test sequences.

Motifs were generated from the *binders* peptide lists using GLAM2, with a starting width of five amino acids, 1,000,000 iterations without improvement, 10 runs, and an alphabet of equal proportions of the 20 amino acids. GLAM2SCAN was used to search the corresponding test sets for sequences matching the motif with the alphabet set as the default protein alphabet for the monoclonal antibodies or equal amino acid frequencies for the polyclonal antibodies. GLAM2SCAN output is the score for each place the motif matches in the test sequence set. The test sequences were ranked by the highest score match within each sequence.

The RELIC Fastaskan program was used to align the binding peptides to the test dataset. The *binders* peptide lists were uploaded as the affinity selected peptides and the corresponding test dataset was uploaded as the FASTA file. Random peptides were not subtracted. Fastaskan compares each five amino acid window of the test sequence with the selected peptide sequence and summing scores of the alignments above a threshold. It outputs a score for each test sequence corresponding to the window of maximum similarity between the peptides and that sequence.

(Table 2)

For both the GLAM2SCAN and the RELIC analysis, the rank of the true epitopes was compared to the test sequences using ROC analysis. A Matlab script to calculate the true positive and false positive rate for each score cutoff was obtained from http://theoval.cmp.uea.ac.uk/~gcc/matlab/roc/ and modified to smooth tied scores. The area under the ROC curve was also calculated using a Matlab script from the same website. The area under the curve will be used to predict the probability of finding an epitope in a database of a

given size. We will assume positive and negative examples will be selected from a database of a fixed size without replacement weighted by the probability that a positive is chosen over a negative as estimated by the area under the curve.

Results. The peptide array consists of randomly generated 17aa sequences with a 3aa, C-terminal linker. The length was chosen for two reasons. Practically, commercial sources of peptide synthesis limit the length to 20aa for large syntheses. Second, we have found that peptides longer than 20aa tend to assume secondary structure and are less soluble (saj, unpublished data). We chose to print 10K peptides because this was the maximum number that could be printed in duplicate on one standard slide. Standard glass slides were used in order to facilitate the processing of the slides on standard equipment and to reduce the cost.

In order to evaluate the peptide microarray platform, examples of antibodies against a known set of variable types of epitopes were chosen. Five monoclonal antibodies with known linear epitopes, and five examples of anti-peptide polyclonal mouse sera raised against peptides selected from the array were used as the test set. Together these epitopes cover a wide range of lengths and physiochemical properties (Table 1). These antibodies will allow us generate a dataset to test how well different sequence analysis approaches are able to predict these antibody epitopes. The monoclonal antibodies were found to bind to a median of 64.1% (range 37.6% - 74.9%) of the random peptides above the slide surface background and secondary only controls. Polyclonal sera showed similar peptide reactivity with a median of 63.6% (range 54.0% - 68.6%). The rank of the peptide used for immunization ranked within the top 100 peptides by signal intensity in four out of the five examples (Fig. 3). Replicate slides had an average Pearson correlation of 0.785 for monoclonals and 0.764 for polyclonals. The heatmap (Fig. 4) shows that

each antibody has a distinct binding pattern on the array. While there is some overlap between the peptides bound by each antibody, about 22% of the top 500 peptides recognized by each antibody are not recognized by the other nine antibodies tested. The uniqueness of the peptides recognized by each antibody implies that the peptide sequences may contain information about antibody specificity.

(Figure 2, 3, 4, and 5)

Each peptide sequence was scored for similarity against each protein sequence. Most of the peptides bound by the antibodies did not show strong sequence similarity to the epitope (Table 2). However, there was some enrichment for sequence similar peptides among the binders. Most of the peptides bound are mimotopes rather than having any obvious similarity to the epitope.

(Table 3)

In order to assess the predictive power of these sequences the alignment of the peptides to the epitopes was compared to their alignment with a set of negative examples. The RELIC alignment program was able to align binding peptides to all of the monoclonal epitopes and 62.7% of the negative examples. The true epitopes had an average score of 14.3 while the negative examples had an average score of 5.9. The ROC analysis found an area under the curve of 0.87 indicating that a true epitope has an 87% chance of having a higher score than randomly selected negative example. All of the polyclonals also had positive peptide alignment scores as well as 86.5% of the positive examples. The true epitopes had an average score of 14.7 while the negative examples had a score of 15.2. The ROC analysis (Fig. 6) indicates that a positive example has a 46% chance of having a higher score than a negative example based on the area

under the curve. The monoclonal epitopes were predicted well by this method, while the polyclonal predictions were similar to chance.

An algorithm capable of detecting subtle patterns may be able to garnish predictive power from these peptide sequences. Convergent motifs were identified for all of the antibodies using GLAM2. The motifs for the monoclonal antibodies ranged from three to five amino acids in width. The polyclonal motifs were four to five amino acids wide. The monoclonal motifs matched the epitope sequences with an average score of 3.5, while the negative examples had an average score of -3.7. Polyclonal motifs matched the immunizing peptide with an average score of 3.8 while the negative examples had an average score of 3.7. The ROC analysis demonstrates that the monoclonals epitopes have an 89.8% chance of being scored higher than the corresponding negative examples in the motif analysis while the polyclonals have a 67.9% chance of scoring higher than the negatives. The motif finding approach demonstrated predictive power on both datasets.

In order to test if combining the two approaches may improve the predictive ability, the scores from the RELIC analysis and the GLAM2 analysis were each scaled to have a minimum score of zero and a maximum score of one and averaged. The ROC analysis was performed on the averaged scores (Figure 6 and 7). The area under the curve was 0.92 for the monoclonals and 0.69 for the polyclonals. Based on the probability estimated from the ROC analysis, there is about a 70% chance of finding a monoclonal epitope in the top ten windows out of a one hundred amino acid protein. There would be a 21% chance of correctly identifying a polyclonal epitope in a small virus among the top 100 hits out of a possible 1000 amino acid database, which is a two-fold enrichment.

(Figure 6 and 7)

While a decoy sequence set was used to estimate the accuracy of these prediction methods, the analysis was also run against the antibody target protein sequences to illustrate how the method would work for predicting an epitope of a known target. P53 was used as the example since both mAb3 and mAb4 are directed against different epitopes of P53, and represent the easiest and hardest epitopes to predict respectively of the five monoclonal examples. Figure 8a. and b. illustrate how well the motif finding and alignment approaches may work when binding peptides show some enrichement for sequence similarity to the epitope, as demostrated by mAb3. When compared against the whole P53 sequence, both approaches predict the true epitope region with the highest score (Figure 8c). In contrast, the true epitope region was not predicted among the top regions by either method for mAb4 (Figure 8d.)

(Figure 8)

Discussion. We have shown that a diverse set of antibodies will each bind a high percentage of the 10K random sequence peptides on our arrays. Each antibody bound a unique set of peptides with little overlap between sets. The list of peptides that bind best to each antibody is only slightly enriched for peptides with sequence similarity to a known epitope or immunogen. However, motif finding and alignment approaches were able to score monoclonal epitopes well among a set of decoy sequences. Predicting the immunogen in a polyclonal response was more difficult, with only the motif finding approach having predictive power.

Antibodies are conventionally characterized into two groups: polyspecific antibodies that recognize many different antigens at low affinity and monospecific antibodies that recognize one or few antigens at high affinity (23-24). Polyspecific antibodies typically have flexible paratopes which allows them to interact with antigens of a variety of shapes and the conformational

entropy lost upon binding prohibits high affinity interactions. In contrast monospecific antibodies tend to be closer to a lock and key binding model, which implies that only antigens with very similar shapes should bind (25-26). Here we found that a set of commonly used and well characterized monoclonal antibodies is capable of binding to many peptides of unrelated sequences when presented on the surface of a microarray. We also observed that polyclonal sera raised to one peptide was able to bind to a median of 20 peptides with higher intensities than the one the sera was raised against, and hundreds at intensities within a fold change of the cognate peptide (Fig. 3).

Weak motifs can still be found using a Gibbs motif sampler algorithm as implemented in GLAM2. The GLAM2 implementation was chosen because it allows for the insertion of variable lengths gaps in the motif. However, none of the motifs found had any insertions or deletions indicating that exact spacing is probably important in binding. Though the patterns identified were subtle, they were repeatedly found over multiple runs. These motifs were used to search a test dataset and were able to score true epitopes higher than decoy sequences. Another approach aligning one sequence at a time using RELIC also showed some predictive power, especially with the monoclonal examples. This indicates that a few sequences with a highly similar window can be informative. The example analysis for the two P53 antibodies illustrate how this data may be used in practice to predict an epitope. The two extremes are depicted where the true epitope is clearly predicted correctly and when it is missed. It is interesting to note that when the motif and alignment methods clearly agree, the correct region is predicted suggesting that the concordance may be reason for confidence in the prediction.

We have identified several sources of error inherent to our microarray such as the presence of impurities and truncations from the peptide synthesis and are currently developing a

new version of the microarray to mitigate these problems. Optimizing the parameters of the motif finding or alignment may be helpful, but would require a larger training dataset to avoid overfitting to these specific examples. The polyclonal is likely a more difficult problem because there may be multiple epitopes within the peptide sequence, but the algorithms are looking for one peak. It may be possible to modify the algorithm to pick up on multiple epitopes in one protein. The biggest limitation of this approach is most likely the sparse sampling of sequence space dictated by the peptide microarray technology employed here. However, much higher density peptide arrays may be produced by in situ synthesis methods. Currently there are oligonucleotide arrays commercially available that have 10^6 features, and it should be possible to achieve similar densities with peptide arrays (27). A peptide library of that size could be designed to cover all possible 5mer sequences, which would enable exact matches to short epitopes to be present.

The prediction accuracy observed in the monoclonal test dataset would be sufficient to narrow down possible linear protein epitopes to a few peptides and have a good chance of correct identification. A much more challenging but interesting question is predicting the immunogenic antigen(s) from polyclonal sera, as would be required based on an immunosignaturing experiment. It will be important for investigators to know the amount of uncertainty present in inferring epitopes from peptide sequences as more of these studies are performed. When identification of the antigen raising the immune response is desired, a number of approaches are possible depending on the biological information available. For example, if one had an immunosignature to a pathogen and wanted to identify the immunogenic epitopes within that pathogen, one could use the prediction methods described here in conjunction with other information available such as subcellular localization to prioritize proteins for further testing. In

a more difficult scenario, one may suspect a pathogen may play a role in a chronic disease and would want to use the immunosignature to identify the pathogen. A more concrete idea of the sequence preference of the sera would be needed in order to search a database of all pathogens than could be inferred directly from the random peptide array results. Alternatively, if technology allowed ready production of new sets of peptides, one could build another array with peptides around the space of the original binders to hone in on the relevant sequence.

Acknowledgements. The authors thank Dr. J. Bart Legutki for providing the mouse sera and John Lainson for microarray production.

References.

1. Fack, F., Hügle-Dörr, B., Song, D. Queitsch, I., Petersen, G., and Bautz, E. K. F. (1997) Epitope mapping by phage display: random versus gene-fragment libraries.

2. Reineke, U. Antibody epitope mapping using de novo generated synthetic Peptide libraries. *Methods Mol. Biol.* 524, 203-11.

3. Irving, M. B., Oscar, P., and Scott, J. K. (2001) Random-peptide libraries and antigen-fragment libraries for epitope mapping and the development of vaccines and diagnostics. *Curr. Opin. Chem. Biol.* 5, 314-324.

4. Wang, L. F. and Yu, M. (2004) Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. *Curr. Drug Targets.* 5, 1-15.

5. Yip, Y. L. and Ward, R. L. (1999) Epitope discovery using monoclonal antibodies and phage peptide libraries. *Comb. Chem. High Throughput Screen.* 2, 125-38.

6. Gershoni, J. M., Roitburd-Berman, A., Siman-Tov, D. D., Tarnovitski Freund, N. and Weiss, Y. (2007) Epitope mapping: the first step in developing epitope-based vaccines. *BioDrugs.* 21, 145-156.

7. Cortese, R., Relici, F., Galfre, G. Luzzago, A., Monaci, P., and Nicosia, A. (1994) Epitope discovery using peptide libraries displayed on phage. *Trends Biotechnol.* 12, 262-7.

8. Bongartz, J., Bruni, N. and Or-Guil, M. (2009) Epitope mapping using randomly generated Peptide libraries. *Methods Mol. Biol. 524, 237-46.*

9. Williams B. A., Diehnelt C. W., Belcher P., Greving M., Woodbury N. W., Johnston S. A., and Chaput J. C. (2009) Creating Protein Affinity Reagents by Combining Peptide Ligands on Synthetic DNA Scaffolds. *J Am Chem Soc*. 131, 17233-41

10. Boltz K. W., Gonzalez-Moa M. J., Stafford P., Johnston S. A., and Svarovsky S. A. (2009) Peptide microarrays for carbohydrate recognition. *Analyst*.134, 650-2.

11. Morales Betanzos C., Gonzalez-Moa M. J., Boltz K. W., Vander Werf B. D., Johnston S. A., and Svarovsky S. A. (2009) Bacterial glycoprofiling by using random sequence peptide microarrays. *Chembiochem*. 10, 877-88.

12. Legutki, J. B., Mitchell, M. D., Stafford, P. and Johnston, S. A. (2010) A General Method for Characterization of Humoral Immunity Induced by a Vaccine or Infection. *Vaccine*. 28, 4529-4537

13. Reineke, U., Ivascu, C., Schlief, M., Landgraf, C., Gericke, S., Zahn, G., Herzel, H., Volkmer-Engert, R. and Schneider-Mergener, J. (2002) Identification of distinct antibody epitopes and mimotopes from a peptide array of 5520 randomly generated sequences. *J. Immunol. Methods.* 267, 37-51.

14. Stephen, C. W., Helminen, P., and Lane, D. P. (1995) Characterisation of Epitopes on Human p53 using Phage-displayed Peptide Libraries: Insights into Antibody–Peptide Interactions. *J. Mol. Biol.* 248, 58-78.

15. Breitling, F. and Little, M. (1986) Carboxy-terminal regions on the surface of tubulin and microtubules. Epitope locations of YOL1/34, DM1A and DM1B. *J Mol Biol.* 189, 367-70.

16. Afonin, P. V., Fokin, A.V., Tsygannik, I.N., Mikhailova, I.Y., Onoprienko, L.V., Mikhaleva, I.I., Ivanov, V.T., Mareeva, T. Y., Nesmeyanov, V. A., Li, N., Pangborn, W. A., Duax, W. L., Pletnev, V. Z. (2001) Crystal structure of an anti-interleukin-2 monoclonal antibody Fab complexed with an antigenic nonapeptide. *Protein Sci.* 10, 1514-21.

17. Rigoutsos, I. and Floratos, A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics.* 14, 55-67.

18. Jonassen, I. (1997) Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.* 13, 509-22.

19. Bailey, T. L., Williams, N., Misleh, C. and Li, W. W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 1, W369-73.

20. Frith, M. C., Saunders, N. F. W., Kobe, B. and Bailey, T. L. (2008) Discovering Sequence Motifs with Arbitrary Insertions and Deletions. *PLOS Comput Biol.* 4, e1000071.

21. Mandava, S., Makowski, L., Devarapalli, S., Uzubell, J. and Rodi, D.J. (2004) RELIC A bioinformatics server for combinatorial peptide analysis and identification of protein-ligand interaction sites. *Proteomics* 4, 1439-1460.

22. Bastas, G., Sompuram, S. R., Pierce, B., Kodela, V., and Bogen, S. A. (2008) Bioinformatic Requirement for Protein Database Searching Using Predicted Epitopes from Disease-associated Antibodies. *Mol. Cell. Proteomics* 7, 247-56.

23. Michaud, G. A., Salcius, M., Zhou, F., Bangham, R., Bonin, J., Guo, H., Snyder, M., Predki, P. F. and Schweitzer, B. I. (2003) Analyzing antibody specificity with whole proteome microarrays. *Nat. Biotechnol.* 21, 1509-12.

24. Zhou, Z. H., Tzioufas, A. G. and Notkins, A. L. (2007) Properties and function of polyreactive antibodies and antigen-binding B-cells. *J Autoimmun.* 29, 219-28.

25. James, L. C., Roversi, P. and Tawfik, D. S. (2003) Antibody multispecificity mediated by conformational diversity. *Science* 299, 1362-7.

26. Mariuzza, R. A. (2006) Multiple paths to multispecificity. *Immunity.* 24, 359-61.

27. Gao X, Pellois JP, Na Y, Kim Y, Gulari E, Zhou X. (2004) High density peptide microarrays. In situ synthesis and applications. *Mol Divers*. 8, 177-87.

Figure Legends

Figure 1. Experimental design schematic.

Figure 2. Array images. Two representative array images are shown in pseudocolor, with zoom in on the same block on each array to the right.

Figure 5. Histograms of Peptide Binding. The distributions of signal intensities for the polyclonal antibodies are shown as histograms with the signal intensities on the y-axis and the relative length of the bar proportional to the number of peptides with that signal intensity. The arrows indicate the location of the peptide that the polyclonal antibody was raised against and the number indicates the rank.

Figure 4. Heatmap of arrays. Data was normalized to the median on each array and clustered using GeneSpring. Averages of replicates are shown. Red represents high signals, yellow intermediate signals and blue low signals

Figure 5. Pie chart of peptide binding showing the overlap between the top five hundred peptides for each of the ten antibodies. Of the 10,340 peptides, 2166 are in the top 500 peptides for at least one of the ten antibodies. 1103 peptides are unique to one of the antibodies and only 6 peptides are recognized by all ten antibodies.

Figure 6. Epitope prediction accuracy. ROC curves of epitope predictions among decoy dataset. The true positive (TP) rate is plotted on the y-axis and the false positive (FP) rate is plotted on

the x-axis. The area under the ROC curve (auroc) indicates the probability that a true epitope would rank higher than a decoy sequence. a. RELIC alignment predictions monoclonal epitopes. b. RELIC alignment predictions of polyclonal epitopes. c. GLAM2 motif predictions of monoclonal epitopes. d. GLAM2 motif predictions of polyclonal epitopes. e. Combined predictions of monoclonal epitopes. f. Combined predictions of polyclonal epitopes.

Figure 7. Chance of finding true epitope in a database. Using the probability calculating from the ROC analysis, the chance of finding an epitope in a database was calculated. The x-axis shows the number of hits that would be examined, and the y-axis shows chance the true epitope would be found among those hits. a. Chance of finding a monoclonal epitope within a dataset of 100 peptides. b. Chance of finding a polyclonal epitope with a set of 1000 peptides.

Figure 8. Example analysis of epitope predictions on P53. The anti-P53 monoclonals mAb3 and mAb4 had the best and worst enrichment respectively for sequence similar peptides among the binders. a. Motif identified by Glam2 analysis of mAb3 binders peptides is similar to mAb3 epitope sequence RHSVV. b. RELIC MATCH alignment of mAb3 binder peptides to epitope region. c. & d. Scores for glam2 motif (- -), RELIC alignment (---), and averaged scores (---) along the p53 sequence for mAb3 (c) and mAb4 (d). The respective epitope locations are highlighted in yellow.

Tables.

Table 1. Antibodies used in this study. mAb1 is a monoclonal antibody raised against a peptide. mAb2 through mAb5 are monoclonal antibodies with epitopes previously determined in the literature. All of the polyclonal antibodies were raised against peptides selected from the random sequence peptide array.

Name	Immunogen	Clonality	Isotype	Clone Name	Sequence	pI	Hydro-pathicity
mAb1	HA peptide	Monoclonal	IgG1	16B12	YPYDVPDYA	3.56	-0.9
mAb2	Human IL2	Monoclonal	IGg1	LNKB2	KPLEEVLNL	4.53	-0.044
mAb3	Human p53	Monoclonal	IgG1	PAb240	RHSVV	9.76	-0.02
mAb4	Human p53	Monoclonal	IgG2b, IgG2a	DO-7, BP53-12	SDLWKL	5.55	-0.25
mAb5	Human Tubulin-alpha	Monoclonal	IgG1kappa	DM1A	AALEKD	4.67	-0.583
pAb1	KLH-peptide	Polyclonal	IgG detected	NA	MDQDDGEGV-IGHFHPILGSC	4.21	-0.185
pAb2	KLH-peptide	Polyclonal	IgG detected	NA	EFWDKEWHTR-ADWPVWDGSC	4.43	-1.245
pAb3	KLH-peptide	Polyclonal	IgG detected	NA	TIPAHNIFWI-LYFSIGTGSC	6.4	0.87
pAb4	KLH-peptide	Polyclonal	IgG detected	NA	PAMKHREPH-WVIPGIIWGSC	8.64	-0.13
pAb5	KLH-peptide	Polyclonal	IgG detected	NA	EFSNPTAQVF-PDFWMSDGSC	3.49	-0.315

Table 2. Comparison of RELIC and GLAM2 approaches

RELIC	GLAM2
Compares peptide sequences and database sequences pair-wise	Looks for a motif within all of the peptide sequences, then uses motif to search database
Uses a five amino acid window	Starts with a five amino acid window, but can adjust window size
Scores include amino acid similarity	Finds patterns using identity only

Table 3. Sequence similarity vs. Binders. The *binders* column indicates the number of peptides selected from the microarray experiments and the *aligners* are the number of peptides that had sequence similarity with the known epitope as described in the Methods section. The number of peptides in common between the binders and aligners lists is in the *both* column. *Expected* column lists the number of peptides expected to be in common if lists were drawn at random from the peptide library. *Ratio* is the ratio of *both* to *expected*. The percentage of the *binders* list that is in *both* is in the last column. The p-value describes the probability that the *binder* and *aligner* list overlap as much as in the *both* list by chance based on the Fisher's Exact test.

	binders	***aligner***	***both***	***expected***	***ratio***	***% of binders align***	***P-value***
mAb1	350	181	**6**	5.88	1.02	1.70%	0.551
mAb2	391	379	**21**	13.75	1.53	5.40%	0.062
mAb3	379	188	**36**	6.61	5.45	9.50%	<0.001
mAb4	354	96	**2**	3.15	0.63	0.60%	0.334
mAb5	369	365	**6**	12.5	0.48	1.60%	0.016
pAb1	258	722	**26**	17.28	1.5	10.10%	0.053
pAb2	258	755	**26**	18.07	1.44	10.10%	0.068
pAb3	263	710	**18**	17.32	1.04	6.80%	0.493
pAb4	274	742	**22**	18.86	1.17	8.00%	0.424
pAb5	267	699	**21**	17.31	1.21	7.90%	0.301
average	316.3	483.7	**18.4**	13.07	1.55	6.20%	0.232

Figure 1.

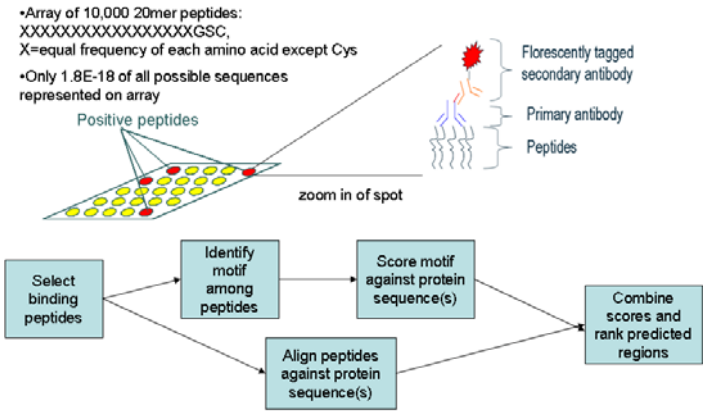

Figure 2.

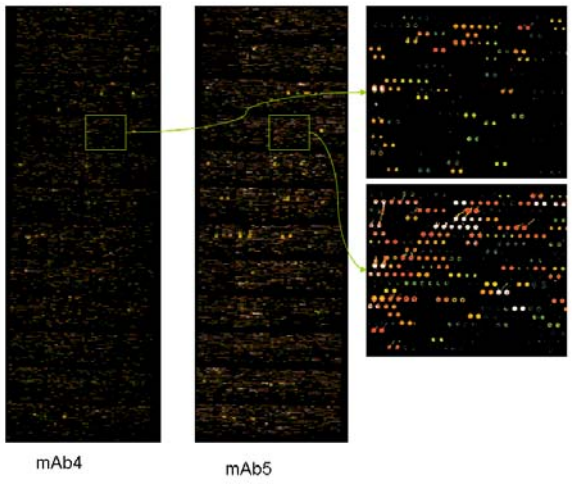

Figure 3.

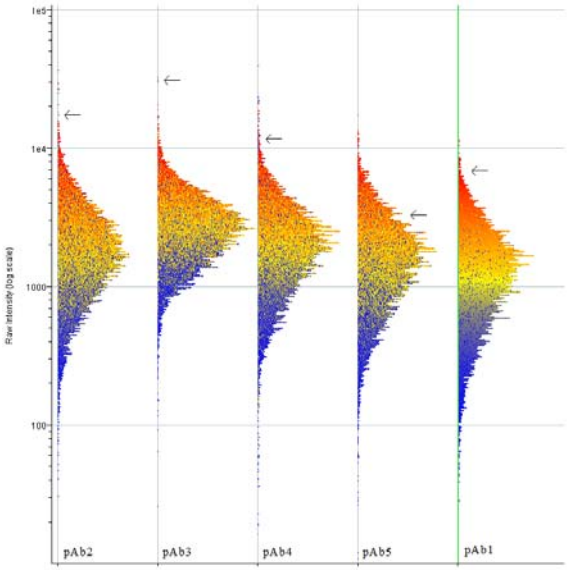

Figure 4.

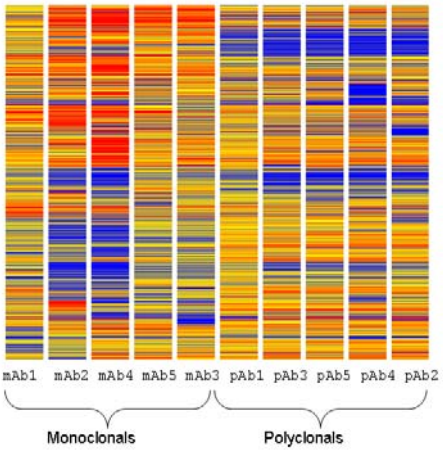

Figure 5.

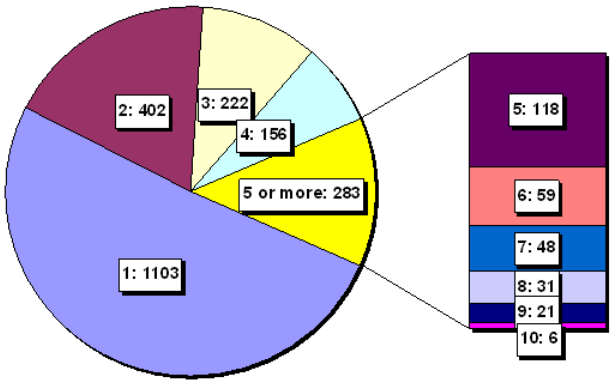

Figure 6.

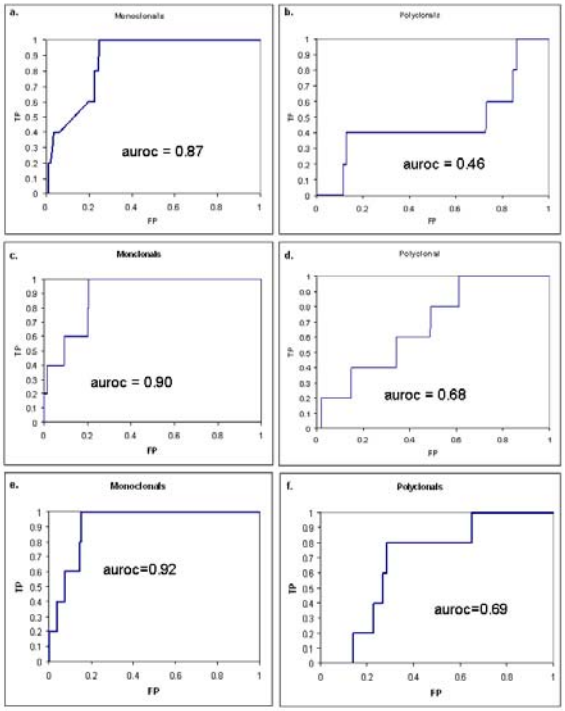

Figure 7.

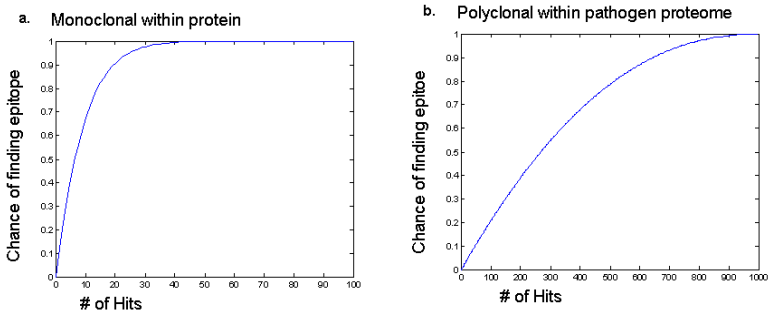

Figure 8.