

Bioinformatic Requirements for Protein Database Searching Using Predicted Epitopes from Disease-associated Antibodies*

Gerassimos Bastas‡, Seshi R. Sompuram‡§, Brian Pierce¶, Kodela Vani§, and Steven A. Bogen‡§||

We describe a new approach to identify proteins involved in disease pathogenesis. The technology, Epitope-Mediated Antigen Prediction (E-MAP), leverages the specificity of patients' immune responses to disease-relevant targets and requires no prior knowledge about the protein. E-MAP links pathologic antibodies of unknown specificity, isolated from patient sera, to their cognate antigens in the protein database. The E-MAP process first involves reconstruction of a predicted epitope using a peptide combinatorial library. We then search the protein database for closely matching amino acid sequences. Previously published attempts to identify unknown antibody targets in this manner have largely been unsuccessful for two reasons: 1) short predicted epitopes yield too many irrelevant matches from a database search and 2) the epitopes may not accurately represent the native antigen with sufficient fidelity. Using an *in silico* model, we demonstrate the critical threshold requirements for epitope length and epitope fidelity. We find that epitopes generally need to have at least seven amino acids, with an overall accuracy of >70% to the native protein, in order to correctly identify the protein in a nonredundant protein database search. We then confirmed these findings experimentally, using the predicted epitopes for four monoclonal antibodies. Since many predicted epitopes often fail to achieve the seven amino acid threshold, we demonstrate the efficacy of paired epitope searches. This is the first systematic analysis of the computational framework to make this approach viable, coupled with experimental validation. *Molecular & Cellular Proteomics* 7:247–256, 2008.

We describe a new platform discovery technology that harnesses the ability of the humoral immune response to identify disease-associated proteins. In many diseases, the immune system generates antibodies to one or more proteins associated with the etiology of the disease. Identifying these proteins could prove valuable in understanding disease

pathophysiology, developing diagnostic reagents, and generating novel therapeutic interventions. However, without clinical clues as to the cellular or microbial source of the protein, it has not previously been possible to identify it. Our Epitope-Mediated Antigen Prediction (E-MAP)¹ technology addresses this need.

Previous investigators have described the mapping of immunodominant epitopes associated with antiviral antibody responses using random peptide combinatorial libraries expressed in phage (1–5). In those cases, the goal was to identify peptides (expressed in a combinatorial phage library) that bind to antiviral antibodies in a similar fashion as the binding of the antibody to the virus itself. Often, consensus amino acid motifs from phage peptide inserts partially identify the epitope, which is then aligned with a span from the antigen's amino acid sequence. It has not previously been possible, however, to use these epitope motifs to identify unknown target proteins from the entire protein database. Epitope reconstruction in this manner has previously yielded data that were insufficiently accurate or informative (6–8). A short motif (4–6 amino acids) does not possess enough information content to uniquely identify a candidate antigen in broad bioinformatic searches. The list of search results retrieved from the nonredundant (*nr*) database is usually large, with hundreds or thousands of hits effectively burying the true matching protein in the noise of extraneous results.

E-MAP has a different clinical context than epitope mapping. With E-MAP, we start with an antibody to an unknown protein antigen. If the antibody is important in a disease-relevant context, then identifying the protein antigen to which it binds will have clinical importance. E-MAP is for identifying proteins that are immunoreactive with an antibody without any prior information as to the identity of that protein. We solve the bioinformatics hurdles through new methods in phage panning and data analysis. E-MAP provides an entirely new capability with which to explore disease pathophysiology. In a separate paper, we describe the application of this

From §Medical Discovery Partners LLC, 715 Albany St., L803, Boston, MA 02118, the ‡Department of Pathology & Laboratory Medicine, Boston University School of Medicine, Boston, MA 02118, and the ¶Bioinformatics Program, Boston University, Boston, MA 02215

Received, March 13, 2007, and in revised form, July 16, 2007.

Published, MCP Papers in Press, September 25, 2007, DOI 10.1074/mcp.M700107-MCP200

¹ The abbreviations used are: E-MAP, epitope-mediated antigen prediction; ER, estrogen receptor; MEME, Multiple EM for Motif Elicitation; MAST, Motif-Alignment and Search Tool; PCR, polymerase chain reaction; PR, progesterone receptor; PSSM, position-specific scoring matrix; TBST, Tris-buffered saline with 0.5% Tween-20.

technology to link multiple myeloma as a malignancy often arising from human herpesvirus 5 (cytomegalovirus)-immuno-reactive lymphocytes (9). The observation was only suspected after E-MAP analysis, and offers insight into the possible etiology of that disease. In this report, we explain the underlying E-MAP methodology. To the extent that the immune system responds to disease-associated proteins, we believe it may have broad clinical applicability.

EXPERIMENTAL PROCEDURES

Phage-display Libraries and Biopanning

Phage libraries contained rationally designed combinatorial libraries of peptide sequences inserted into the N' terminus of the cplII minor coat protein of the M13 bacteriophage. The libraries were supplied by Dyax Corp. (Cambridge, MA). The libraries termed TN6 and TN10 contained two conserved cysteine residues separated respectively by four or eight amino acids. The cysteines formed a disulfide bridge, creating a conformationally constrained ring (10). Trinucleotide-mutagenesis technology, involving controlled polymerization of preformed trinucleotides, was used to diversify the amino acids within the ring and three amino acids on either side of the ring, allowing all amino acid types (except cysteine) with equal frequency (11). The libraries have $\sim 1 \times 10^9$ independent transformants, a measure of library diversity.

The libraries were screened by biopanning using standard methods (12, 13) with a few modifications. Briefly, paramagnetic beads coated with anti-mouse IgG (Dynabeads; Dynal Corp., New York) were prepared by mixing either the estrogen receptor (ER)- or progesterone receptor (PR)-specific mouse mAbs (for positive enrichment) or the polyclonal mouse IgG (for negative depletion) and incubating overnight at 4 °C on a rotator. Antibody-adsorbed Dynabeads were washed five times with phosphate-buffered saline (PBS) containing 0.05% Tween-20 and twice with PBS before use in biopanning of phage libraries. A TN6 or TN10 phage library containing 10^{11} – 10^{12} plaque-forming units was negatively depleted by incubation with Dynabeads (100 μ l) coated with polyclonal mouse IgG for 1 h at room temperature on a rotator. This negative depletion step removes phage that may bind to constant regions of mouse IgG. The unbound phage (supernatant) were then positively selected on the (ER or PR-specific) target mAb-adsorbed Dynabeads. The phage library was incubated with the mAb-coated beads for 2–3 h on a rotator. The beads were washed 10 times with PBS containing 0.05% Tween-20 and three times with PBS to remove nonspecifically bound phage. Phage particles that bound to the mAb-coated beads were eluted with 0.1 mol/L glycine-HCl (pH 2.2) containing 1 g/L bovine serum albumin. The recovered eluate was neutralized with 1 mol/L Tris-HCl (pH 9.0). To ensure that the bound phage were completely eluted, the beads were treated a second time with elution buffer and the eluate was neutralized. The two eluates were pooled. The eluted phage were amplified and used in a second round of biopanning. After two rounds of positive selection, *Escherichia coli* were infected with the cultured phage and grown on agar plates.

Post-panning Selection of Phage Clones

Replicate plaque lifts were created by laying nitrocellulose membranes onto the aforementioned agar plates, at 4 °C for 1 h. The membranes were marked for orientation, carefully lifted from the agar and placed at 65 °C to dry for 5 min. The membranes were blocked with 5% milk in Tris-buffered saline with 0.5% Tween-20 (TBST) and washed twice with TBST. The selecting (ER- or PR-specific) mAb was prepared in TBST (2.5 mg/L) and placed on the membrane for 2 h at room temperature or at 4 °C overnight. The membranes were then

washed eight times with TBST and incubated with anti-mouse-IgG-HRP (Sigma, St. Louis, MO) (1:5000 dilution) for 1½ hours. A chemiluminescence protocol was used to visualize patterns of immunoreactivity (ECL Western blotting Detection Reagents, Amersham Biosciences). Developed films could be oriented onto the corresponding agar plates. The most immunoreactive spots (representing distinct plaque colonies) were picked and grown individually for further analysis. A second replicate lift was usually obtained and worked up in like manner as a control, testing nonspecific immunoreactivity of the phage clones to mouse polyclonal IgG.

DNA Insert Sequencing

Phage clones that had high specific immunoreactivity for the selecting antibody were submitted for further analysis, by sequencing the nucleotide inserts coding for the combinatorial peptides. The sequencing template was prepared by polymerase chain reaction (PCR) amplification from an overnight phage culture. The primers used for PCR were 5-CGGCGCAACTATCGGTATCAAGCTG-3 and 5-CATGTACCGTAACACTGAGTTTCGTC-3. Thirty rounds of PCR were performed on an MJ Research Tetrad thermocycler (MJ Research, Inc.). The PCR product was diluted 1:20 with distilled H₂O. Sequencing was performed in both the forward and reverse directions with the following primers: 5-GATAAACCGATACAATTAAAGGCTCC-3 and 5-GTTTGTCTGCTTTCCAGACGTTAG-3. ABI Big Dye™ (version 1.0) was used to perform a 5- μ l sequencing reaction [2 μ l of Big Dye, 1 μ l of distilled H₂O, 0.5 μ l of primer (at 3 pmol/ μ l), and 1.5 μ l of diluted PCR product]. The samples were then cycled for 45 rounds on an MJ Research Tetrad thermocycler. After cycling, 2.5 volumes of absolute ethanol were added, and the mixture was centrifuged at 1850 $\times g$ for 30 min. The plates were inverted over paper towels, and then centrifuged at 100 $\times g$ for 30 min. The samples were resuspended in 5 μ l of distilled H₂O and detected on an ABI 3700 DNA Analyzer.

Motif Elicitation and Bioinformatic Search of the Protein Databases

The determined nucleotide sequences of the inserts were translated *in silico* using the Translate tool from ExPASy Proteomics Server of the Swiss Institute of Bioinformatics web utility available at ca.expasy.org. The translated protein sequences could be verified to be in frame by identification of invariant elements of the cplII protein and the hallmark presence of the invariant cysteines. The variable regions of the inserts were transcribed into the FASTA form and submitted to MEME (Multiple EM for Motif Elicitation, available at meme.sdsc.edu/meme). The MEME output contains the submitted peptides rank-ordered for the presence of the dominant motif determinants. Separate files were created containing the position specific scoring matrix (PSSM) characterizing the predicted motif.

To carry out bioinformatic searches using a single motif, the PSSM was submitted to the MAST (Motif-Alignment and Search Tool) utility (meme.sdsc.edu/meme), to be searched against the *nr* protein database while allowing a maximal E value (expectation value). The first 500 hits were then screened for the presence of the known target. For pairwise motif searches, the PSSMs from two motifs were combined and submitted to MAST. The MAST database search program will return many hits, which can be ranked by their position *p* value, sequence *p* value, and combined *p* value of alignment. These terms are defined, and the program more thoroughly described, at meme.sdsc.edu/meme/mast-output.html. Briefly, when tentative matches are found, each is given a score, reflecting how well the motif's PSSM fits the particular span from the identified sequence. The position *p* value of an alignment is defined as the probability of a random span in a

randomly generated sequence having a match score at least as large as that of the given motif. The sequence itself is assigned a p value that is defined as the probability of a random sequence of the same length having a match score at least as large as the highest scoring match in the sequence. MAST also assigns a combined p value, defined as the probability of a randomly generated same length sequence having sequence p values whose product is at least as small as that of the matches of the motifs to the given sequence. Based on the latter determination, an expectation value (E value) is generated by multiplying the combined p value of a sequence by the number of database entries. The E value can then be thought to represent the expected number of sequences in a random database of equal size that would match the motif(s) at least as well.

For most of our analyses, we set the E value to <10 and the threshold value for motif display to $p \leq 0.0001$ (default setting). For our analyses, the p value threshold of 0.0001 is relatively lax, as the correct matches had p values several log orders lower. The size of the displayed list of database matches is therefore effectively gated by the E value. Setting the E value higher will result in a longer list of matches, in decreasing order of amino acid sequence homology. Otherwise, it does not change the rank order of database matches. Any proteins found with a qualifying E value of <10 solely on the basis of a single motif were disqualified, as described in the Results. All possible pairwise combinations of the four determined motifs' PSSMs were analyzed in this manner.

Modeling the Bioinformatic Requirements of Predicted Epitopes as Search Probes

Sequence Generation—To generate sets of sequences for MEME/MAST analysis, short sequences of predefined length N were selected randomly from the NCBI *nr* protein sequence database. These sequences were then used to construct a position specific probability matrix, with the degree of residue conservation at each position perturbed by a Gaussian function around the average conservation, C . These matrices were used to generate 20 “pseudo-epitopes” (mock phage clone peptide inserts), also termed “pseudo-clones.” The pseudo-clones contained the epitope motif at random positions within a 20-mer, flanked by randomly generated residues. Therefore, these pseudo-clones contained combinatorially scrambled motifs, each with varying degrees of sequence conservation relative to the chosen native protein epitope sequence, but on the whole approaching the defined average conservation when looked at as a group.

Single Motif Searches—For each target epitope, sequences were generated as described above. These pseudo-clone sequences were used as an input to the motif searching tool MEME. Parameters for MEME included use of the zoops model and user-defined restriction of the motif length. The MEME output motif was then given to MAST, which was used to search the *nr* database. Success was defined as recovering the original protein sequence within the top 10 MAST database hits. The above-described test was performed 40 times for each value of N and C . These success rates were averaged over 40 runs to obtain an average and standard deviation.

Multiple Motif Searches—To generate the success rates for two motifs in a pairwise search, proteins were randomly selected from the *nr* database and random spans were chosen as target epitope sequences. For each protein, two nonoverlapping epitopes of lengths 5–8 amino acids were randomly chosen from the *nr* database. Each epitope was used to generate pseudo-clones (as described above) which were then processed with MEME. Both MEME motifs were then given to MAST. The average success rate and standard deviation were calculated as for the single motif searches. For pairwise searches, we set the MAST threshold expectation value (E value) to

≤ 10 and the threshold value for motif display to $p \leq 0.0001$. This effectively returns hits that have high scoring alignments for both motifs. A few searches retrieved hits with a qualifying E value that was based solely on the match to one epitope, without a corresponding second epitope match. These search results were disqualified, as they are contrary to the requirement of pairwise search strategy, that both epitopes must be present in the candidate protein.

RESULTS

E-MAP Protocol Overview

The E-MAP method incorporates two components, illustrated schematically in Fig. 1. First, we use a random peptide combinatorial library to elucidate the amino acid sequence of the antibody's epitope. We also refer to this elucidated peptide epitope as the “predicted” epitope, in that it is predicted by amino acid sequence analysis of the peptide inserts from antibody-binding phage clones. The predicted epitope is a consensus motif, revealing which amino acids are most likely present at each position. Fig. 1 illustrates two hypothetical peptide epitopes derived from a single protein. The amino acids are arbitrarily designated with the letters A-F and L-Q.

To identify meaningful protein matches from predicted epitopes, it is important to maximize the certainty about the identity of each amino acid in the sequence. Uncertainty in the predicted epitope can inappropriately skew the content of the retrieved hit list. Using peptide phage display (14), we are essentially carrying out a casting process on a molecular scale. We are filling the antibody's binding site (the “paratope”) with random oligopeptides, and identifying which peptide sequences are the highest affinity binders. We then reconstruct a virtual best fitting consensus motif by analyzing the commonalities of those peptide sequences. We use the MEME (Multiple Expectation-maximization for Motif Elicitation) software utility to identify consensus motifs in the sequenced peptide inserts (15, 16). MEME creates a consensus motif profile, capturing each phage clone's sequence information in PSSM, a two-dimensional numeric array. The profile is, in essence, a virtual mimotopic array of the peptides that bind to the antigen-binding site of the antibody. Using such a profile in a bioinformatic search offers distinct advantages. Instead of searching with a single “best-guess” query representing the dominant motif, the queried profile considers a larger number of combinatorially weighted sequences, averaging around the dominant motif.

We usually find certain positions in a motif to be invariant while others may exhibit conserved substitutions. These substitutions generate uncertainty in knowing the sequence of the native protein, affecting the size of the database search hit list and potentially skewing its contents. In our experience, a clear consensus motif usually emerges if high stringency screening techniques (“Experimental Procedures”) are used during the phage display component.

The second step in the E-MAP process (Fig. 1) is the bioinformatic search of the protein database using the predicted epitope as an *in silico* probe. From our computer model

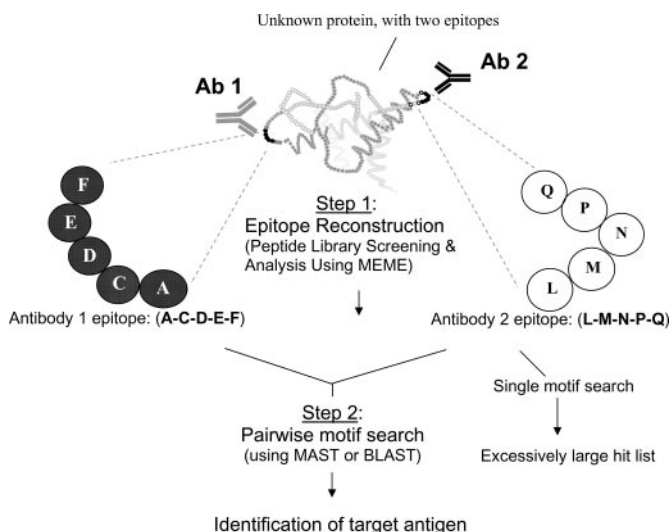


FIG. 1. Schematic representation of the two-step process comprising the E-MAP technology. Two antibodies, labeled “Ab1” and “Ab2” are directed to two different linear epitopes on a hypothetical protein antigen. These epitopes are in *bold* on the protein antigen and also shown in an exploded view. The identity of the amino acids is arbitrarily designated with the letters A-E or L-P, for illustrative purposes. In step 2, the predicted epitopes, identified by phage display of peptide combinatorial libraries, are used in pairwise submissions to search the protein databases.

(Fig. 2) and experience, individual motifs are generally not sufficiently long to yield accurate searches from the *nr* protein database. The combined statistical power of a pairwise search, however, is sufficient to narrow the list to a small number of antigen candidates. We use the MAST (Motif Alignment and Search Tool) utility (16, 17) to perform single and pairwise motif searches against the *nr* protein database. The pairwise submission finds proteins containing both predicted epitopes. The retrieved hits are ranked according to their combined E value, which evaluates the two epitopes’ degree of alignment to the database entry (18).

Bioinformatic Requirements for the Consensus Motif Profile

There are two important variables for identifying proteins from only short linear epitopes: 1) the length of the epitope and 2) the fidelity with which the predicted epitope matches the actual sequence in the protein database (average motif conservation). We expected that longer epitope lengths (more information) and higher average motif conservation (greater accuracy) will both increase the likelihood of obtaining a correct database match. However, the thresholds for each parameter are unknown. For the process to yield accurate database matches, how long and accurate must the predicted epitope be?

We analyzed the relationship of epitope length and average motif conservation on the success rate in protein database searching with an *in silico* simulated analysis. Short stretches of proteins, randomly selected from the *nr* protein database,

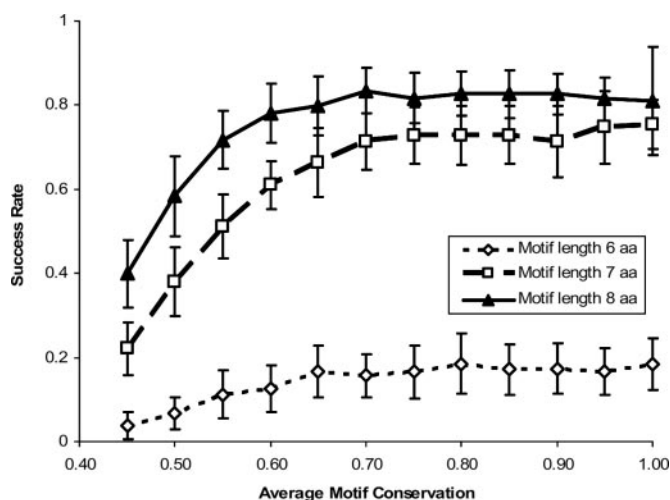


FIG. 2. The peptide epitope length and average motif conservation are important variables in searching the protein database. The average motif conservation is the proportion of homologous amino acids between each pseudo-clone and the corresponding actual native sequence. The “success rate” is defined as the proportion of protein database searches that resulted in the correctly matching protein among the top 10 database hits. Each point represents the mean $\pm$ S.D. of 40 searches from 40 different randomly selected proteins.

were randomly selected as putative antibody epitopes. We then simulated peptides from a combinatorial phage library with varying degrees of homology to the randomly chosen epitopes. We termed each of these simulated peptides a “pseudo-clone” because the peptide sequence was computationally generated and not actually derived from a phage clone peptide insert. The amino acid sequences of the pseudo-clones were then run through the MEME and MAST bioinformatic algorithms and scored for the predicted epitope’s ability to identify the target protein (described in “Experimental Procedures”). Fig. 2 represents the output from the computer simulation, demonstrating the inter-relationship of epitope length and motif conservation. In Fig. 2, the average motif conservation (x-axis) is the proportion of amino acids that are identical between each pseudo-clone and the corresponding actual native sequence. The “success rate” (y-axis in Fig. 2) is the frequency with which the correct match showed up among the top 10 protein database search results.

Fig. 2 illustrates that predicted peptide epitopes with length ≥ 7 amino acids begin to have enough information so as to be capable of potentially yielding correct hits by single epitope searching. There is a significant difference in the predictive capability between a 6-mer and 7-mer peptide; 7 amino acids appears to be a threshold value.

Fig. 2 also illustrates that, for 7-mer and 8-mer motifs, an average motif conservation of 0.6 yields successful matches in more than half of the searches. An average motif conservation of ~ 0.6 – 0.7 appears to be another threshold value. The success rate drops precipitously with lower average motif

conservation numbers. It should be remembered that average motif conservation in Fig. 2 refers to the degree of homology for each individual pseudo-clone, not the final consensus motif. The overall accuracy is higher when many pseudo-clones are used to collectively search the database, as a PSSM. Errors at one amino acid position for a particular pseudo-clone tend to be neutralized by the other pseudo-clones, which, on average, are accurate at that position. Thus, the consensus motif is nearly 100% accurate even when the average motif conservation for each individual peptide (x-axis) is only 60–70%, accounting for the graph's plateau.

In our experience, epitope reconstruction by phage display of peptide combinatorial libraries typically yields a consensus motif that is four to six amino acids long. Based on the results illustrated in Fig. 2, we would expect it to be too short to be useful for protein database searching. Hundreds of irrelevant close matches effectively bury and oftentimes exclude the true match from the viewable retrieved hit list. Although most short predicted epitopes have insufficient information content to yield accurate hits on their own, they can still be highly predictive in the context of pairwise searching. Pairwise analysis can accurately identify protein matches from the protein database that were otherwise too low in the priority ranking based on a single epitope search.

Pairwise Epitope Search Concept

Pairwise epitope submissions to the protein database dramatically increase the statistical power of a search, beyond what is possible with a single epitope. Querying two motifs simultaneously asks which proteins contain both predicted epitopes. From a clinical standpoint, it requires that there are two or more antibodies to a target protein antigen, both of which will provide information about the protein's identity. In practice, one often cannot be certain that pairs of antibodies from patients are, in fact, directed to the same target. This problem can be surmounted, as described later.

We tested this hypothesis *in silico*, measuring the success rate for a pairwise submission strategy. The details of the computer modeling are similar to those used for Fig. 2, except that we searched with two predicted epitopes instead of only one. Both epitopes were present on the same protein. The average motif conservation was held constant at 0.7, a typical figure in our hands. The results of pairwise searching using our computer model are listed in Table I. Unlike single motif submission, the combination of two motifs now becomes highly predictive (67–88% success rate with 5- or 6-mer peptides). This success rate is in contrast to the expected result if each motif were searched individually ($\leq 15\%$ success rate).

In Vitro Validation: Overview

We tested these theoretical computer predictions in an *in vitro* model system, using monoclonal antibodies to the ste-

TABLE I
Likelihood of success with pairwise submission

Motif 2 length	Motif 1 length (number of amino acids)			
	5	6	7	8
5	0.67 $\pm$ 0.07	0.84 $\pm$ 0.05	0.87 $\pm$ 0.05	0.89 $\pm$ 0.05
6		0.88 $\pm$ 0.05	0.88 $\pm$ 0.05	0.89 $\pm$ 0.05
7			0.89 $\pm$ 0.05	0.90 $\pm$ 0.05
8				0.91 $\pm$ 0.05

roid hormone receptors human ER and PR. We treated the binding specificities of the antibodies (human ER and PR) as unknowns for the purpose of this study. Our goal was to determine if we could identify the antigens solely on the basis of the predicted epitope sequence data and bioinformatic analysis.

Monoclonal antibodies designated "1" and "3" bind to the human ER, whereas antibodies 2 and 4 bind to the human PR. These particular four monoclonal antibodies were chosen because they were already in the lab and well characterized. We have no reason to believe that the results would be materially different had we chosen alternative antibodies.

To obtain the highest average motif conservation, we learned that it is best to use high stringency screening methods. For example, average motif conservation improved if the enriched (second round) phage clones were then further screened by immunoblot (Fig. 3). We selected the most immunoreactive phage clones by creating plaque lifts and immunoblotting, probing the blot with the monoclonal antibody chosen for positive selection (Fig. 3). Sequencing the peptide inserts of only the most immunoreactive second round phage clones results in greater concordance and accuracy in defining the consensus motif sequence. Table II shows the peptide sequences that were entered into MEME and the consensus motifs at the top.

In Vitro Validation: Epitope Reconstruction of Four Exemplary Antibodies

Because of the stringent phage panning selection process, the individual phage peptide inserts had a high degree of consensus. The average positional conservation of each motif ranged from 73.25% to 95.2%. Even though there was a high degree of homology among the individual peptides, the derived consensus sequence is not always an exact match to the native epitope.

Antibody 1—The consensus motif is SR(S/G)CXSY. The corresponding sequence in the native protein is ARSPRSY. Antibody 2—The consensus motif is QAPYY, which is a close match to the native sequence QVPYY. Alanine (A) and valine (V) are conserved substitutions.

Antibody 3—The consensus motif is GDF(P/S)DCAY, similar to the native linear sequence of GDFPDCAAY. In this case, the invariant cysteine forced the selection of phage

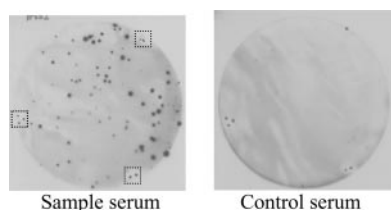


FIG. 3. **Representative immunoblot of enriched phage.** The phage library was panned against monoclonal antibody-coupled paramagnetic beads. The *left-hand blot* represents the image after immunodetection using the relevant monoclonal antibody. The *right-hand blot* represents the image after immunodetection with a (negative) control antibody. The boxes identify spots that we placed on the replicate lifts for purposes of alignment.

clones containing the relevant peptides anchored around its position.

Antibody 4—The predicted sequence, LHQCQ, was close to the native sequence LHQIQ. The difference is due to the invariant cysteine (C) being substituted for isoleucine (I).

In each case, monoclonal antibody specificity was confirmed by testing for immunoreactivity to a synthetic peptide containing the epitope (19). With these predicted epitopes in hand, we then asked if we could have deduced the correct protein from a protein database search using single or pairwise searches.

Identification of Antigens from the Protein Database

Single Motif Searches—Single motif searches are not generally successful, unless the epitope length is unusually long (e.g. eight amino acids). In the single motif submission analysis against the *nr* database, the heptamer SR(S/G)CXS_Y (monoclonal antibody 1, PR-specific) was unable to find PR in the first 500 hits (not shown). QAPYY (monoclonal antibody 2, ER-specific) also failed to retrieve the correct protein in the top 500 hits. The pentamer LHQCQ (monoclonal antibody 4, ER-specific) retrieved the human estrogen receptor in positions 40 and 43, too low in the list to establish identification.

The only exception was the octamer GDF(P/S)DCAY (monoclonal antibody 3, PR-specific). A search of the *nr* database identified PR and PR homologues as the top-ranked hits. The fact that an 8-mer is able to identify the correct protein in a single motif search agrees with the predicted 7 amino acid requirement described in the context of Fig. 2. However, obtaining a long (8-mer) predicted epitope with such a high degree of sequence fidelity to the native protein is unusual. To better reflect a more typical, shorter, predicted epitope, we arbitrarily shortened the octamer to a hexamer by removing the two C-terminal amino acids. With a (now shortened) predicted epitope of GDF(P/S)DC, a markedly different hit list results. Human PR is at position 26 and below.

Pairwise Motif Searches—The outcomes of database searches for single *versus* pairwise submissions are markedly

different. Unlike single motif searching, pairwise searches identify the true protein target from the *nr* protein database. Table III shows that the pairwise submission of Antibodies 1 and 3 (PR-specific antibodies) correctly identified PR. For antibody 3, we used the shorter hexamer predicted epitope that we previously described rather than the complete octamer, so as to make the analysis more realistic. The pairwise submission for antibodies 2 and 4 (ER-specific antibodies) also correctly identified ER.

Mismatched epitopes, from different proteins, can generally be distinguished. When working with antibodies to unknown protein antigens, it is generally not possible to know, *a priori*, if they target the same protein. For pairwise epitope searching to be practical, mismatched pairs should not yield database search results that will mislead a research investigation. Namely, we want two linear epitopes on the same protein to converge in identifying that protein out of the protein database. However, it is possible that one or both of the antibodies actually binds to a determinant that is dependent upon the three-dimensional conformation of the target protein. In performing our analysis, we might then identify a linear amino acid sequence that approximates the conformational epitope. That predicted epitope derived from phage display does not actually exist as a linear sequence on the target protein, but it might inadvertently match an irrelevant protein in the *nr* protein database. This might potentially lead to spurious database matches. We wanted to determine how serious this potential pitfall is and if there is a way to avoid it.

We predicted that these mismatched epitopes will not likely be problematic for E-MAP analysis. It is reasonably likely that a conformation-dependent epitope, represented as a linear amino acid sequence, will match other (irrelevant) protein database entries. Almost any 5- or 6-mer amino acid sequence will have reasonably close matches in the protein database. However, pairwise searching requires that both epitopes match linear sequences contained within the same protein database entry. The likelihood that two conformation-dependent epitopes, both represented as linear amino acid sequences, will have matches to the same protein database entry seems quite low. We tested this prediction by running mismatched database searches, shown in Table IV.

We found that mismatched pairwise epitope searches can usually be distinguished from correctly paired searches. Predicted epitopes that do not belong together generally yield few search results. The hits that do result usually show less amino acid sequence identity and more conserved substitutions. Table IV shows the search results of four mismatched pairs of predicted epitopes. Inappropriately paired predicted epitopes result when the two antibodies are directed to different antigens, in this case between epitopes for the human ER and PR. The same situation would exist if one of the antibodies binds to a conformational determinant since such epitopes are not represented in the protein database. Table IV shows that there are a few database hits with these inappro-

TABLE II
Consensus motifs of four exemplary monoclonal antibodies

Antibody 1 (PR) SRSCxSY (G)		Antibody 2 (ER) QAPYY		Antibody 4 (ER) LHQCQ	
1-24	GYCQDATSLLSRGCKSY	2-47	TQGCDQAPYYIGRCDHP	4-54	DRACFLHQCQDY
1-32	PLCMDSGSKMSRSCCKSY	2-43	QNRCDQAPYYGHPECFWH	4-53	NFQCYLHQCQFP
1-16	SRSCCKSYPEQYQRTCLF	2-2	PDSCNSQAPYYMSCYGD	4-52	PSMCFHQCQGI
1-28	PVCLKPRENRSRGCSY	2-51	GDXCDFQAPYFHQCHPP	4-51	KNICDLHQCQLT
1-12	SRGCASYSTIYDQKCIN	2-33	SSICSQAPYFDQHCVFEE	4-50	AQHCFHQCQSL
1-56	ARGCHSYLESYNRKCSH	2-19	DKTCEANYQAPYICDST	4-45	HFNCFLHQCQPF
1-57	SQCRFDGEQASRGCRSY	2-50	VFGCKPQAPYVTGCDHV	4-44	KGNCFLHQCQIE
1-52	SRGCLSYGKLYFADCTS	2-42	WEWCTQAPYWHAGCAQL	4-43	GSLCFHQCQNT
1-2	SRCAAEAVSLSRGCRSY	2-30	PFNCQTPYYTDHGCMKS	4-41	YSSCFHQCQLS
1-45	FRCHNEHYGESRGCSY	2-38	QRGCQAPFYQDGXCTTA	4-39	KNDCYLHQCQHY
1-43	SRSCHSYLSQCMPP	2-34	VEECDRQAPWYQVCTLE	4-38	FDECFHQCQPR
1-19	YNCKYRFDNQSRSCHSY	2-37	TTKCNTPSPNQAPYCRTH	4-60	TEMGLHQCQHP
1-9	SRSCASYMGWCPGG	2-49	SXNCVXGWQAPYXCIQM	4-24	SGHAKLHQCQGR
1-62	SRACKSYGIWCEMP	2-4	SNLCSHQAPYPSLCNRG	4-26	KLHWQHTCDTL
1-53	ARGCQSYWMADEFSCYG	2-44	MEDCDQVPYLLPHCESH	4-21	KLHWQYECCHAW
1-31	PTCADYTDVWARSCHSY	2-3	HRSCQTPTPYHTMDCDYW	4-23	RLHQYQGDGCMQ
1-46	HSCHMDDNNHSRSCRSY	2-48	HSQCNGQRPYLGNCDEVQ	4-20	LHQTKHTCKKH
1-30	SRSCLSYEDKQIAPCSF	2-45	SRDCGNDWFAPYVCGDP		
1-20	RTCTHPQVNNSRSCRSY	2-29	KMHCFQKPYWDEVCDVN		
1-4	SRSCISYGFHQSGMCEL	2-35	KRDCMFENQPPYFCPPQ		
1-3	WLCDFPSFTEESRSCRSY	2-1	RFHCDSMLDQTPYCMPL		
1-33	PACFWQQHGISRSCCKSF	2-5	KDACDNNGYQWPYCMGK		
1-7	SRSCMSYWEIPPECGG				
1-6	SRSCQSYFEHQVMGCPL				
1-44	SVCPPNTMNM SRGCSSY				
1-26	ARSCRSYAPEEXTACDP				
1-11	HLCMHDKKEFGSRSCGSY				
1-5	SRGCMSEFFGPMAETCPP				
1-1	SRSCSSYYSQHNMCSG				
1-29	SRSCASFQAMSMACVP				
1-34	DLCEMDYGKARSCHSY				
1-27	SRSCXSYNXMNFWCPR				
1-17	SRSCRSFPTWDNNQCQV				
1-13	SRSCFSYRTFPSAXCTN				
1-55	SRSCRTYNSHDSPYCRS				
1-51	DFCADFDQSN SRSCYSY				
1-47	SRACSSYFQEEYAICVS				
1-2	SRSCSTFHHGQASPVCF				
1-21	SSCNIFEHFEGRGCSY				
1-8	GRGCSYYGQNTDRCNR				
1-14	SRSCSSFHRQDRKCLP				
		Antibody 3 (PR) GDFPDCAY			
		3-5	GDFSDCAYAQQEQLVYQC		
		3-3	GDFSDCAYSSGPWHICKP		
		3-2	GDFHDCAYVLGSKQDPCY		
		3-1	GDFPDCVYPMQHSLQCK		
		3-4	GDFDDCAYTLGWDLTCT		
		3-6	GDFPDCFYPLSENKQICH		
		3-7	GDFPXCXYTPMXXASDXX		

privately paired predicted epitope searches when using an E value threshold of 10, the same as used in Table III.

Further analysis reveals that even these few database hits are distinguishable from true matches. So far, we have two threshold criteria: the presence of both motifs in the candidate matching protein and a low E value (e.g. ≤ 10 in the examples shown.). In analyzing the database search results in Table IV,

we found additional criteria to help distinguish false from true matches. Namely, false matches tend to have more conserved substitutions and fewer identical amino acid matches. The search algorithm gives partial credit for conserved substitutions, accounting for a low E value. Identifying this difference requires direct visual examination, comparing the search results to the predicted epitopes. In our data set, true matches

TABLE III

Validation: search results for correctly matched pairs ($E < 10$)

Ab1 (PR) x Ab3 (PR)	E value
dbj BAC11012.1 delta4 + 6/2 progesterone r . . .	2.1 ^a
dbj BAC06585.1 Progesterone receptor [Ho . . .	2.2 ^a
dbj BAC11011.1 delta 3 + 6/2 progesterone . . .	2.6 ^a
dbj BAC11013.1 delta 6/2 progesterone re . . .	2.9 ^a
dbj BAB91074.1 delta 4 progesterone rece . . .	3 ^a
gb AAS00096.1 progesterone receptor [Ho . . .	3.8 ^a
ref NP_000917.2 progesterone receptor [Ho . . .	3.8 ^a
sp P06401 PRGR_HUMAN Progesterone receptor	3.8 ^a
ref XP_694039.1 PREDICTED: similar to ann . . .	5.1
ref XP_905331.1 PREDICTED: hypothetical p . . .	8.5
gb AAA29925.1 Sm50 antigen	9.7
Ab 2 (ER) x Ab 4 (ER)	
gb AAW69860.1 estrogen receptor alpha m . . .	2 ^a
emb CAA66784.1 estrogen receptor [Camelu . . .	3.9 ^a
prf 2113258A estrogen receptor	4.7 ^a
gb AAB88052.1 estrogen receptor [Sus sc . . .	5 ^a
gb AAG41358.1 estrogen receptor alpha [. . .	5.1 ^a
ref XP_518813.1 PREDICTED: similar to est . . .	6.4 ^a
gb AAA58462.1 estrogen receptor-related . . .	10 ^a

^a Match that represents the correct protein or protein homologue.

TABLE IV

Validation: search results for incorrectly matched pairs

	E value
Ab 1 (PR) x Ab 2 (ER)	
ref NP_071550.1 CXC chemokine LIX [Rattus . . .	7.6
Ab 2 (ER) x Ab 3 (PR)	
ref ZP_00438202.1 COG3522: Uncharacterized . . .	7.6
Ab 3 (PR) x Ab 4 (ER) (no hits retrieved)	
Ab 1 (PR) x Ab 4 (ER)	
gb EAA43338.1 ENSANGP00000023167 [Anoph . . .	7.5
gb ABB60986.1 hypothetical protein SDY_ . . .	9.5
ref XP_543640.2 PREDICTED: similar to ker . . .	9.9

have the following characteristics, distinguishing them from false ones:

1. For a five amino acid predicted epitope, an identical match in four of five positions.
2. For a seven amino acid predicted epitope, identity in four positions and homology in at least two more.
3. For an eight amino acid epitope, identity in six positions and homology in at least one more.

The threshold criteria for percent identity and conserved substitutions for any motif will probably vary from search to search, depending upon the circumstance. We do not expect these exact thresholds to apply to other data sets, but they provide guidance in correctly prioritizing potential database matches for further analysis. Ultimately, proof of a correct match depends upon *in vitro* testing, demonstrating that the antibody actually binds to the candidate protein.

DISCUSSION

The E-MAP technology represents a valuable new investigative tool for uncovering the target of immune responses. E-MAP has potential applicability in many disease contexts, including lymphoproliferative disorders, inflammatory diseases, allergy, and autoimmunity. In a real-world application of E-MAP technology, we applied it to the investigation of multiple myeloma. In that study, we determined that in approximately one-third of multiple myeloma cases, the malignant cells arise from human herpesvirus 5-immunoreactive lymphocytes (9). This finding is not the first time that an infectious agent is linked to a B lymphoproliferative disorder and raises new possible avenues of therapeutic intervention.

An important new feature of the E-MAP technology is the pairwise search analysis. This feature overcomes the statistical limitation that previously precluded finding accurate matches with most predicted epitopes. Searching the protein databases simultaneously with two, even short, predicted epitopes provides sufficient statistical power to accurately identify the correct target. This pairwise analysis can yield strikingly different results compared with single search protocols currently in use. Top ranking hits from single epitope searches are usually incorrect since even a single amino acid substitution can dramatically skew the search results. Because of this potential for error, top ranking search results from single epitope database searches may exhibit complete sequence identity in their alignment with the predicted epitope probes and still be incorrect matches. Indeed, dozens or even hundreds of database hits may be exact matches or have only one amino acid substitution, depending upon the length of the predicted epitope. Pairwise motif analysis, on the other hand, combines the predictive power of two motifs, thereby establishing a higher level of search stringency. The net result is the reorganization of candidate hit lists compared with single epitope searches, revealing a new set of search results with the requisite presence of both motifs appearing in declining order of relative combined alignment.

A second important requirement for E-MAP is the use of high stringency biopanning methods. When used properly, the predicted epitope is nearly identical to the eliciting epitope in the native antigen. This is a testament to the power of the phage display technique that provides an antibody with a staggering array of oligopeptides from which to select. By imposing high stringency selection conditions, proper phage to antibody ratios, and a post-panning immunoblot selection of individual clones, the selected phage clones' peptide inserts generally observe a tight convergence to the native protein epitope. There is always some degree of uncertainty in predicting epitopes using phage-displayed combinatorial peptide libraries. We have shown, however, that a small amount of uncertainty can be tolerated in the bioinformatic search algorithm.

Retrieving a protein database match does not prove that

the antibody is actually capable of binding to that protein. It is important to separately test the antibody for its ability to bind to the protein in question. The statistical tools available through the MAST program, such as the p and E values, primarily affect the length of the retrieved database match list. MAST now uses a default p value of 0.0001, which is relatively nonstringent for our purposes. All of the close matches have p values lower than .0001, and they are usually several log orders lower. Consequently, the E value is mainly responsible for determining whether a potential match will be displayed or not. Once a list of close matches is obtained, we then visually inspect it to determine if the low E value is due to both pairwise submissions or just one, how much of the match is exact *versus* conserved substitutions, and does the candidate match make sense in the context of the disease.

A limitation of E-MAP is that conformational epitopes will not yield matches in the protein database. Although some textbooks suggest that conformational epitopes may predominate in immune responses, we think that this conclusion may somewhat overestimate their prevalence. Many antigens also produce humoral immune responses to linear epitopes (20). Indeed, we previously described that the monoclonal antibodies used for clinical immunohistochemistry testing are all directed to linear epitopes (19). The search tools that are currently available for epitope mapping of conformational epitopes require knowledge of the crystal structure of the protein antigen (21). Although predicted epitopes derived from antibodies to conformational epitopes are not helpful in identifying the protein target, Table IV demonstrates that they also will likely not create many false leads.

In practical terms, E-MAP involves submitting a collection of clinically relevant monoclonal antibodies for analysis not knowing which, if any, are correctly matched to the same protein. Since we have no way to know which antibody pairs will be correctly matched, we submit all combinations in separate pairwise searches. The number of independent pairwise combinations to be performed is, in fact, manageable and calculated from combination theory, as $n!/[2 \times (n-2)!]$, where n equals the number of independent antibodies being analyzed. For example, nine different antibodies results in 36 different pairwise searches.

Conclusion—We describe a new technique termed E-MAP that can provide important clues to disease etiology and offering new capabilities in drug discovery. E-MAP is a useful investigative tool in clinical situations where there is reason to believe that an immune response may be associated with, or identify, the cause of a disease. E-MAP results do not independently prove that a particular protein is an antibody's target. Rather, E-MAP identifies a short list of potential protein candidates for further testing and evaluation. By applying this common sense approach, E-MAP promises to be valuable for investigating a broad array of diseases. We have recently applied E-MAP for the analysis of multiple myeloma paraproteins

and found a surprisingly high incidence of herpesvirus immunoreactivity, especially to human herpesvirus 5 (9).

Acknowledgment—We are grateful for the helpful discussions with Prof. Zhiping Weng, Boston University, Boston, MA.

* This study was supported by National Institutes of Health Grants CA94557 and CA106847. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

||To whom correspondence should be addressed. Tel.: 617-638-4103; FAX: 617-638-4103; E-mail: sbogen@bu.edu.

REFERENCES

- Enshell-Seijffers, D., Smelyanski, L., Vardinon, N., Yust, I., and Gershoni, J. (2001) Dissection of the humoral immune response toward an immunodominant epitope of HIV: a model for the analysis of antibody diversity in HIV+ individuals. *FASEB J.* **15**, 2112–2120
- Folgori, A., Tafi, R., Meola, A., Felici, F., Galfre, G., Cortese, R., Monaci, P., and Nicosia, A. (1994) A general strategy to identify mimotopes of pathological antigens using only random peptide libraries and human sera. *EMBO J.* **13**, 2236–2243
- Scala, G., Chen, X., Liu, W., Noel Telles, J., Cohen, O., Vaccarezza, M., Igarashi, T., and Fauci, A. (1999) Selection of HIV-specific immunogenic epitopes by screening random peptide libraries with HIV-1 positive sera. *J. Immunol.* **162**, 6155–6161
- Prezzi, C., Nuzzo, M., Meola, A., Delmastro, P., Galfre, G., Cortese, R., Nicosia, A., and Monaci, P. (1996) Selection of antigenic and immunogenic mimics of hepatitis C virus using sera from patients. *J. Immunol.* **156**, 4504–4513
- Wang, L., and Yu, M. (2004) Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. *Current Drug Targets* **5**, 1–15
- Szecs, P. B., Riise, E., Roslund, L. B., Engberg, J., Turesson, I., Buhl, L., and Schafer-Nielsen, C. (1999) Identification of patient-specific peptides for detection of M-proteins and myeloma cells. *Br. J. Haematol.* **107**, 357–364
- Dybwad, A., Lambin, P., Sioud, M., and Zouali, M. (2003) Probing the specificity of human myeloma proteins with a random peptide phage library. *Scand J. Immunol.* **57**, 583–590
- Zonder, J., Tainsky, M., Stone, M., Konduri, K., Oliver, J., and Ratner, S. (2005) Myeloma paraprotein epitope identification using a phage library of sterically constrained peptides [Abstract]. *American Society of Clinical Oncology Annual Meeting, Orlando, FL*
- Sompuram, S., Bastas, G., Vani, K., and Bogen, S. (2008) Multiple myeloma is commonly derived from HCMV-immunoreactive lymphocytes. *Blood* **111**, 302–308
- McLafferty, M. A., Kent, R. B., Ladner, R. C., and Markland, W. (1993) M13 bacteriophage displaying disulfide-constrained microproteins. *Gene* **128**, 29–36
- Virnekas, B., Ge, L., Pluckthun, A., Schneider, K. C., Wellenhofer, G., and Moroney, S. E. (1994) Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Res.* **22**, 5600–5607
- Smith, G., and Petrenko, V. (1997) Phage display. *Chem. Rev.* **97**, 391–410
- Sparks, A., Adey, N., Cwirla, S., and Kay, B. (1996) Screening phage-displayed random peptide libraries. *Phage Display of Peptides and Proteins*, Academic Press, New York
- Smith, G. P., and Petrenko, V. A. (1997) Phage display. *Chem. Rev.* **97**, 391–410
- Bailey, T., and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28–36, AAAI Press, Menlo Park, CA
- Bailey, T. L., Baker, M. E., and Elkan, C. P. (1997) An artificial intelligence approach to motif discovery in protein sequences: application to steroid dehydrogenases. *J. Steroid Biochem. Mol. Biol.* **62**, 29–44
- Bailey, T., and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54

18. Bailey, T. L., and Gribskov, M. (1998) Methods and statistics for combining motif match scores. *J. Comput. Biol.* **5**, 211–221
19. Sompuram, S. R., Vani, K., Hafer, L. J., and Bogen, S. A. (2006) Antibodies immunoreactive with formalin-fixed tissue antigens recognize linear protein epitopes. *Am. J. Clin. Pathol.* **125**, 82–90
20. Atassi, M. Z. (1984) Antigenic structures of proteins: their determination has revealed important aspects of immune recognition and generated strategies for synthetic mimicking of protein binding sites. *Eur. J. Biochem.* **145**, 1–20
21. Schreiber, A., Humbert, M., Benz, A., and Dietrich, U. (2005) 3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins. *J. Comput. Chem.* **26**, 879–887