Revealing protein structures: A new method for mapping antibody epitopes

Brendan M. Mumey
Department of Computer Science
Montana State University
Bozeman, MT 59717-3880
mumey@cs.montana.edu

Brian W. Bailey
NIH/NIAAA/DICBR/LMBB
Fluorescence Studies
Park 5 Building
12420 Parklawn Dr. MSC 8115
Bethesda, MD 20892-8115
bbailey@mail.nih.gov

Edward A. Dratz
Department of Chemistry and Biochemistry
Montana State University
Bozeman, MT 59717-3400
dratz@chemistry.montana.edu

ABSTRACT

A recent idea for determining the three-dimensional structure of a protein uses antibody recognition of surface structure and random peptide libraries to map antibody epitope combining sites. Antibodies that bind to the surface of the protein of interest can be used as "witnesses" to report the structure of the protein as follows: Proteins are composed of linear polypeptide chains that come together in complex spatial folding patterns to create the native protein structures and these folded structures form the binding sites for the antibodies. Short amino acid *probe* sequences, which bind to the active region of each antibody, can be selected from random sequence peptide libraries. These probe sequences can often be aligned to discontinuous regions of the one-dimensional *target sequence* of a protein. Such alignments indicate how pieces of the protein sequence must be folded together in space and thus provide valuable long-range constraints for solving the overall 3-D structure. This new approach is applicable to the very large number of proteins that are refractory to current approaches to structure determination and has the advantage of requiring very small amounts of the target protein. The binding site of an antibody is a surface, not just a linear sequence, so the epitope mapping alignment problem is outside the scope of classical string alignment algorithms, such as Smith-Waterman. We formalize the alignment problem that is at the heart of this new approach, prove that the epitope mapping alignment problem is NP-complete, and give some initial results using a branch-and-bound algorithm to map two real-life cases.

1. ANTIBODY EPITOPE MAPPING

Proteins are nano-machines that carry out most of the processes in living cells. These tiny machines are constructed from long-chains (typically 100-1000 elements) composed of twenty different amino acids arranged in characteristic sequences. Proteins must be folded into complex 3-D shapes to create the binding pockets and active sites necessary to carry out their myriad of different functions [5]. There are at least 30,000 different proteins in human cells [9] and each protein has a unique folded structure. Whenever the 3-D folding structure of linear protein sequences can be determined this information has provided fundamental insights into mechanisms of action that are often extremely useful in drug design. Traditional methods of protein structure determination require preparation of large amounts of protein in functional form, which often may not be feasible. Attempts are then made to grow 3-D crystals of the proteins of interest for structure determination by x-ray diffraction, however, obtaining crystals of sufficient quality is still an art and may not be possible [24, 25]. Alternatively, if the proteins are not too large, are highly water soluble, and meet other criteria, methods of nuclear magnetic resonance can be used for structure determination [8]. It is also possible to predict 3-D structures *de novo* from the sequence of amino acids in the protein, but the available methods are not very accurate unless a 3-D structure of a highly homologous protein is already known [2] (also see http://predictioncenter.llnl.gov).

A large fraction of protein structures of interest (50% or more) cannot be solved by the traditional approaches above [14, 15]. Thus, the *antibody imprint* method is being developed to provide structural information on difficult cases that appear refractory to traditional approaches [7, 21, 1]. The antibody imprint method makes use of information carried in the structures of antibodies against proteins of interest to reveal the 3-D folding of target proteins [7, 21, 1, 13, 17]. Antibodies tend to be highly specific for the protein structures that they recognize [20]. They can either recognize *continuous* epitopes or *discontinuous* epitopes. Discontinuous epitopes provide the most useful structural information in antibody imprinting, because they can reveal distant pieces of primary sequence that are located in close spatial proximity in the native, folded protein. Most antibodies recognize discontinuous epitopes on protein surfaces [26]. Studies of a substantial number of antibody-protein complexes with known x-ray structures indicate that these complexes form in a *lock* and key manner with little or no structural change induced by complex formation [1 1]. Fortunately, relatively few long-distance constraints are needed to reveal the global folding of proteins [10, 12]. In addition, the spatial proximity of different regions of proteins can change during function and antibody imprinting has the potential to reveal these structural changes, if appropriate antibodies can be found that recognize the different structural shapes [1].

Briefly, the antibody imprinting method is carried out by immobi-

RECOMB '02, **April 18-21, 2002 Washington,** D.C., **USA**
 ISBN **I-58113-498-3-02/04** . ..$5.00

lizing antibodies (against a target of interest) on beads or in plastic wells. Random peptide libraries are exposed to the immobilized antibodies so that library members that bind to the antibodies can be captured on the surface. The random peptide libraries are carried on bacteriophage (which is called "phage display" of the library) [3]. Each phage has a different peptide expressed on the surface of one of the coat proteins of the phage and there are typically 5 . 10^9 [6] and even up to 10^{12} different sequences in the library [29]. These probe libraries contain either linear peptides or can be constrained with circular topology where the two ends of the probe are chemically linked with a disulfide bond. Peptide sequences that do not stick to the antibody are washed off and the tightly binding phage are eluted under harsher conditions. The phages that bind to the antibody are multiplied by growth in suitable bacteria and again exposed to the immobilized antibody. These cycles of binding and enrichment of members of the random peptide library are usually repeated three times to select the phages with the highest affinity to the antibody. These enriched phages are then highly diluted and grown as clones that arise from individual phage particles. Each of the phage clones carry the DNA sequence that codes for the peptide sequences that have been selected. This DNA region is amplified by PCR, tagged with fluorescent primers and sequenced in a standard automated DNA sequencer. In this way, the sequence for each peptide is discovered. These individual sequences are often highly conserved and SO-100 independent peptide sequences together describe a consensus sequence, called the *consensus* epitope of the antibody. The problem addressed in the present paper is to develop a means to examine and evaluate all possible ways in which a consensus epitope can be mapped onto the target protein in question, to provide proximity constraints on the 3-D structure of the protein. We adopt the terminology that the consensus epitope sequence forms a probe that is to be aligned to the protein *target* sequence.

In the remainder of the paper, we formalize the probe-target alignment problem, describe a branch-and-bound algorithm to find optimal (and sub-optimal) alignments, prove the corresponding decision problem is NP-complete, and provide some experimental results for two biologically significant proteins. Finally, we comment on some future directions for this work.

2. FORMALIZING THE PROBLEM

The core idea of the antibody imprint method is that a probe that binds to the active region of a particular antibody is expected to be highly homologous to the binding site of a protein that also binds to the same antibody. We thus are faced with the problem of aligning the probe amino acid sequence, s, to one or more regions of the target protein amino acid sequence, t. Typically, s is about 8-20 amino acids long and t is several hundred. Unlike traditional string alignment problems, we allow for localized sequence rearrangements. This captures the possibility that several loops of the linear protein sequence may be pinched together (possibly with twists) to form the binding site. Additionally, it is possible for local rearrangements of amino acids to occur, reflecting the fact that the binding site of an antibody is a surface, not just a linear sequence. As such, the problem is outside the scope of classical string alignment algorithms such as Smith-Waterman [31]. In general, we will allow any permutation of the probe sequence to align to the underlying protein sequence'. Furthermore, gaps will be permitted in both probe and target sequences. Large gaps can occur in probe sequence when the epitope mapping is discontinuous. We also allow single position gaps in the target, reflecting the possibility of a single residue insertion into the probe. To be a valid alignment, each probe position and target position can be used at most once per mapping. Formally, an alignment A consists of a sorted set $P_A = \{i_1 < i_2 < \ldots < i_k\}$, and another set $T_A = \{j_1, j_2, \ldots, j_k\}$, with the interpretation that the i_p-th probe residue, $s(i_p)$, is aligned to the j_p-th target residue, $t(j_p)$, for $1 \leq p \leq k$.

We adopt a two-part scoring system to evaluate the quality of alignments. The scoring system is composed of a substitution score and a gap cost,

$$\text{score(A)} = S(A) - G(A).$$

The $S(A)$ component is calculated with a substitution matrix M, similar in principal to a Dayhoff matrix, used in other protein alignment contexts. We discuss our choice of substitution matrix in the experimental results section. The matrix is also used to score unaligned probe positions; if character c occurs and unaligned probe position i, then $M(c, -)$ is included in $S(A)$:

$$S(A) = \sum_{p=1}^{k} M(s(i_p), t(j_p)) + \sum_{\text{probe positions } i \notin P_A} M(s(i), -)$$

The probe gap cost $G(A)$ is calculated by examining the number of amino acid residues skipped along the target protein sequence between successive aligned probe positions:

$$G(A) = \sum_{p=1}^{k-1} d[|j_{p+1} - j_p|]$$

where d(z) is the cost of skipping x amino acids along the target between successive mapped probe positions. For circular probes we also include the term $d[|j_k - j_1|]$ in the above sum. The computational problem is thus to find finding an alignment A that maximizes score(A).

A branch-and-bound algorithm can be used to solve this alignment problem in practice. The algorithm constructs a search tree to find the optimal alignment(s). Often, a user may also be interested in near-optimal solutions so the algorithm is designed to find the top r solutions where r is user-specified. Each node in the search tree represents a partial alignment of the probe to the protein sequence. At the root, all probe positions are unaligned. Nodes at level $i > 0$ in the tree fix the alignment of the i-th probe position (either to an available target position or to a "-", indicating an unmatched probe position). A leaf is reached when all probe positions have been considered and each leaf represents a particular alignment. Whenever a new node n is created, an upper bound on the highest possible alignment score achievable in the subtree rooted at n is computed. If this bound is less than the r-th best solution found so far, we can immediately prune the node from the search. Nodes that are on the boundary of current search tree are said to be on the *frontier*. For each frontier node n, an expected score is calculated by dividing n's current score by its depth in the tree. A heap data structure is used to extract a node with maximal expected score from the frontier. This node is then expanded, descendant child nodes are created for each possible alignment of the next probe position. When a leaf is reached, the score of the associated alignment is calculated. This score is compared to the current r-th best solution and if greater replaces it. When such a replacement occurs, the frontier is scanned to cull out any other nodes that can now be eliminated. This algorithm has been implemented as a C++ program called FINDMAP.

'In some cases, e.g. membrane proteins, it may be known that certain regions of the target protein are inaccessible to antibodies and so can be excluded from consideration as potential alignment positions.

The experimental results section presents some of our initial experience with FINDMAP; most problems of interest run in a few minutes or less on a fast workstation.

3. PROBLEM COMPLEXITY

In this section we show that the probe-target sequence alignment problem is NP-complete [16]. We first define a decision version of the problem:

The ALIGN decision problem.
Input: A probe string s, a target string t (over a common alphabet), a substitution score matrix M, a distance penalty function d, an objective score Q.
Output: A decision on whether there exists an alignment with score at least Q.

LEMMA 1. *ALIGN is NP-complete.*

Proof. First note that ALIGN belongs to NP because the score of a given alignment can be checked in polynomial time. We will show that ALIGN is complete for NP via a polynomial time reduction from 3SAT. Consider an instance of 3SAT I_{3S} consisting of a collection of clauses $C = \{c_1, c_2, \ldots, c_m\}$ on a finite set of variables $U = \{x_1, \ldots, x_k\}$. We will describe a polynomial time reduction to an instance $I_A = (s, t, M, d, Q)$ of ALIGN such that a truth assignment exists for U that satisfies C if and only if an alignment between s and t with score at most Q can be found. We construct I_A as follows: The string alphabet used is

$$A = U \cup \{\neg x_1 \ldots \neg x_k\} \cup \{y_1, \ldots, y_k\} \cup \{`c_1', \ldots, `c_m'\} \cup \{`\#', `*', `@'\}.$$

All entries of M are set to $-\infty$ except the following: $M(\alpha, c_i) = 0$ if α is a literal in clause c_i, $M(x_i, y_i) = M(\neg x_i, y_i) = 0$ for all $1 \le i \le n$, and $M(\cdot, `*') = 0$ (here . represents any symbol). For each literal α, let $[\alpha]$ be the multiplicity of a among all clauses in C. The probe string used is

$$s = @\ B_1\ B_2\ \cdots\ B_k$$

where

$$B_i = \underbrace{x_i \ldots x_i}_{[x_i]+1 \text{ copies}}\ @\ \underbrace{\neg x_i \ldots \neg x_i}_{[\neg x_i]+1 \text{ copies}}\ @.$$

Let $n = |s| - (m + k)$. The target string used is

$$t = \underbrace{* * \cdots *}_{n \text{ copies}}\ \underbrace{\# \# \cdots \#}_{n \text{ copies}}\ c_1\ c_2 \cdots c_m\ y_1\ y_2 \cdots y_k.$$

The distance penalty function used is

$$d(l) = \begin{cases} 0 & \text{if } l < n \\ 1 & \text{otherwise.} \end{cases}$$

Observe that $m + k < n$, so only jumps across the central gap of #'s, referred to as the ***bridge,*** will contribute to the gap cost. The leading @ of s forces any finite-score alignment to begin on the left side of the bridge. Note that every non-# letter in the target must be matched in order to completely align the probe (all probe positions must be matched as $M(\cdot, -) = -\infty$). In order to match all of the y_i's, at least one literal from each B_i must be used. Thus each B_i contributes at least one return jump across the bridge. If a literal is matched against a clause symbol c_i, then any truth assignment that makes this literal true will satisfy c_i. We choose $Q = -2k$ to insist that each B_i contributes ***exactly*** one return jump across the bridge. Because the positive and negative literals in each block B_i are separated by an @, only literals of a single polarity can be matched to symbols to the right of the bridge. This ensures a consistent truth assignment. Thus, any alignment with score exactly $-2k$ will produce a satisfying assignment for I_{3S} and vice versa.

4. EXPERIMENTAL RESULTS

In this section, we discuss our initial experimental results using FINDMAP, our implementation of the branch-and-bound alignment algorithm previously described. We discuss two cases: a validation case where the 3-D structure is known and a second case where the structure has not been fully solved. FINDMAP requires an amino acid substitution scoring matrix. We chose the matrix shown in Figure 1, since a very similar substitution matrix was developed by Bordo and Argos [4] for scoring substitutions of protein residues exposed to the aqueous surface. Antibody binding sites on target proteins must be exposed to the aqueous surface for antibody accessibility and so an aqueous-exposed substitution seems appropriate.

Recently, Jesaitis and co-workers studied an antibody against the ubiquitous cytoskeletal protein, ***actin*** [21]. They reported the manual mapping of consensus peptides derived from phage display library selection, to complex epitopes on the surface of actin. The phage-display-discovered peptides could be mapped onto the actin surface to mimic a discontinuous epitope that was consistent with the known 3-D x-ray structure of actin [22]. Figure 2 shows the mapping of one of the consensus sequences, VPHPTWMR, onto the surface of actin. It should be emphasized that this manual mapping utilized knowledge of the actin x-ray structure. The FINDMAP alignment used only the protein primary sequence. The single difference from the manual mapping is FINDMAP's selection of the more buried but plausible Thr 103 (dark green residue) instead of the more exposed Thr 358 (maroon-colored residue) for the T in VPHPTWMR. We viewed this result as a test case for the antibody imprinting technique, validating the process on a known protein structure. It's possible that including an estimate for the probability of surface exposure in the overall alignment scoring function could help in cases such as this [19].

We also used the actin test case to optimize the gap cost parameters. We chose a simple linear penalty function up to a maximum gap penalty that does not further penalize long gaps (long gaps are expected in discontinuous epitopes):

$$d(n) = \min(a \cdot n, b)$$

To search for suitable values of a and b, we ran FINDMAP on the actin example, where the 3-D structure is known, using the probe sequence VPHPTWMR. We tested 140 different combinations of a, b pairs as shown in Figure 3. The deviation from the proper mapping with parameter values that were non-optimal were systematic. When a was set too small, the highest scoring epitopes found were implausibly discontinuous with identity matches widely spread in the mappings at the expense of any allowable amino acid substitutions. In contrast, when a was too large, excessively continuous local epitopes were found, which may include large numbers of very non-favorable amino acid substitutions. In Figure 3, the best parameter choices yielded 18 alignments that had identical optimal scores, of which one agreed exactly with the manual mapping except at one residue position (a reason for the proliferation of optimal solutions is the freedom of the final R to align to a number of positions in the target). We picked a = 0.5 and b = 1.5 from the

target residue

probe residue	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	-
A	1	0	0	0	0	0.5	0	0	0.5	0	0	0	0.5	0	0	0.5	0.25	0	0	0	-1
C		1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1
D			1	0.5	0	0	0	0	0	0	0	0.5	0	0.25	0	0	0	0	0	0	-1
E				1	0	0	0	0	0	0	0	0.25	0	0.5	0	0	0	0	0	0	-1
F					1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.5	-1
G						1	0	0	0	0	0	0.5	0.5	0	0	0.25	0.25	0	0	0	-1
H							1	0	0.25	0	0	0	0	0	0.25	0	0	0	0	0	-1
I								1	0	0.5	0.25	0	0	0	0	0	0	0.5	0	0	-1
K									1	0	0	0	0	0.5	0.5	0	0.5	0	0	0	-1
L										1	0.25	0	0	0	0	0	0	0.5	0	0	-1
M											1	0	0	0	0	0	0	0.25	0	0	-1
N												1	0	0.25	0	0	0	0	0	0	-1
P													1	0	0	0	0	0	0	0	-1
Q														1	0	0	0	0	0	0	-1
R															1	0	0	0	0	0	-1
S																1	0.5	0	0	0	-1
T																	1	0	0	0	-1
V																		1	0	0	-1
W																			1	0.25	-1
Y																				1	-1

Figure 1: Amino acid substitution scoring matrix used in FINDMAP. This matrix is based on the probability of amino acid substitutions on surface-exposed residues of proteins. The Bordo and Argos substitution matrix was modified so that Gly/Pro substitutions score 0.50, Arg/His, Lys/His, and Gly/Ser substitutions score 0.25. Unaligned probe positions were charged a penalty of -1.

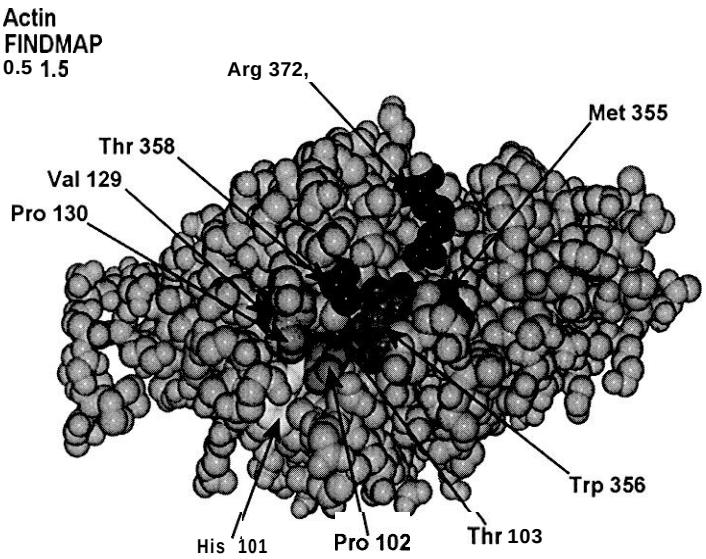
green region for the second experiment discussed next.

The second case we considered was the integral membrane protein *rhodopsin*, the structure of which is not fully known. Rhodopsin is the photoreceptor in dim light vision in retinal rod cells and is an archetype for the structure and mechanism of a large superfamily of cellular receptor proteins that respond to a wide range of hormones and neurotransmitters [33, 23]. The x-ray crystal structure of the dark-adapted, resting structure of rhodopsin was recently published [27, 32] but some of the surface features of the protein were poorly ordered in the crystals and not visible in the x-ray structure. A computational model of the missing surface was built and energy minimized but the structure was uncertain in this region [l]. Antibody imprinting was applied to the surface. One of the antibodies (B 1gN) maps to the extracellular surface of rhodopsin in a compact patch that shows the proximity of two distant segments of sequence, in excellent agreement with a well defined region of the x-ray structure [1] (data not shown). One of the other antibodies studied (4B4) targeted part of rhodopsin where the x-ray structure was not fully resolved, shown in Figure 4A. The single optimal mapping of the 4B4 epitope found with FINDMAP had a discontinuity where the epitope runs from the residue colored red in Figure 4B to orange, yellow and light green. The dark green residue is predicted to be located adjacent to the light green residue by FINDMAP. This is evidence that the surface loop folding shown in Figure 4A and is incorrect and should be adjusted to form a hairpin turn, as shown in Figure 4B. Thus, the antibody imprinting technique appears to be capable of providing new structural information. Many additional antibodies will be required to reveal the complete surface structure of rhodopsin and its light-excited conformations. More detailed antibody imprinting studies on rhodopsin are in progress [1].

5. FUTURE WORK

The antibody imprinting approach described appears to be very general and we are applying it to a number of additional cases. First, we are investigating a number of antibodies where the 3-D structure of the antibody-target protein complex is known to atomic resolution by x-ray diffraction. In these validation cases the correct antibody epitope mappings are known. We plan to use these test cases to refine the substitution matrix and assess the generality of the optimum values of the gap penalty parameters. We are also experimenting with different methods to include consideration of predicted aqueous exposure in FINDMAP. Second, we are applying the antibody imprinting technique to several integral membrane proteins that are difficult structural targets and to reveal the nature of functional conformational changes in membrane proteins [1].

It has not escaped our attention that all the steps in the antibody imprinting process are adaptable to high throughput enhancement. As more experience is gained with this technique, especially with the known test cases to aid in refining the substitution matrix, we plan to add high throughput enhancements to the epitope selection, epitope sequencing, and epitope mapping. In some cases suitable antibodies are already available, but in many cases suitable antibodies have not been prepared or do not provide sufficient coverage of the surface or the conformational states of interest. In the absence of available antibodies, the rate-limiting step in the current process is the isolation and characterization of new antibodies. Technology to express random antibody libraries on phage [28, 18] has recently been developed and will allow much more rapid identification of specific antibodies. Affinity maturation steps [28] applied to antibodies selected from random libraries are also adaptable to high throughput approaches. The random antibody libraries are uniquely useful for rapid selection against transient protein conformations, which are expected to reveal important information on protein mechanisms. An interesting computational problem arises in the context of using a random antibody library: given a set of a probe-target alignments, can they be clustered into groups corresponding to unique epitopes of the target protein? This is a natural question as the identities of the random antibodies that bind the target protein will not be known in general. A probe-target alignment can be found for each probe found that binds at least one of these




```
DEDETTALVCDNGSGLVKA########################################################## 1-70

                                345                             12
                                HPT                             VP
####################YNELRVAPEEHPTLLTEAPLNPKANREKMTQIMFETFNVPAMYVAIQAVL 71-140

SLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRLDLAGRDLTDYL###################### 141-210

########################################################################## 211-280

#####################################################APPERKYSVWIGGSILASLS 281-350

    76              8
    MW              R
TFQQMWITKQEYDEAGPSIVHR 351-372
```

Figure 2: Mapping of **epitope VPHPTWMR** onto the surface of **actin** manually, and by **FINDMAP.** The independent manual mapping required the knowledge of the **actin** x-ray structure. The top-scoring **FINDMAP** alignment having the best match is shown **(#'s** indicate previously known non-binding regions of **actin).** Mapped residues are color coded in rainbow order from red to purple for the **FINDMAP** results, based on the probe **peptide** sequence from N to C-terminal. The manual and **FINDMAP** mappings differ only in their alignment of Thr **358** (maroon) where **FINDMAP** tends to pick Thr 103 (dark green).

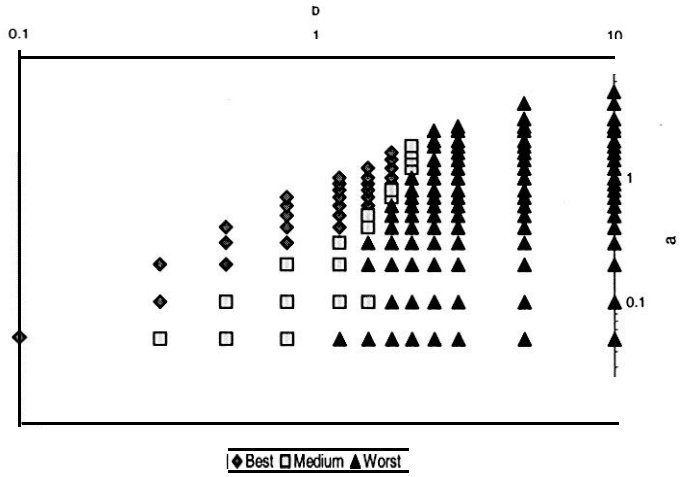


Figure 3: FINDMAP target gap penalty parameter sensitivity. For the distance penalty function $d(n) = \min(a\,n, b)$, various combinations of a and b where tested on mapping the probe epitope VPHPTWMR to the actin sequence (see Figure 2). These alignments were ranked into three categories based on how closely they agreed to the published manual mapping on the known 3-D structure of actin. The green points in the figure indicate parameter combinations where FINDMAP found the published mapping to within one residue position as one of the top-scoring alignments.

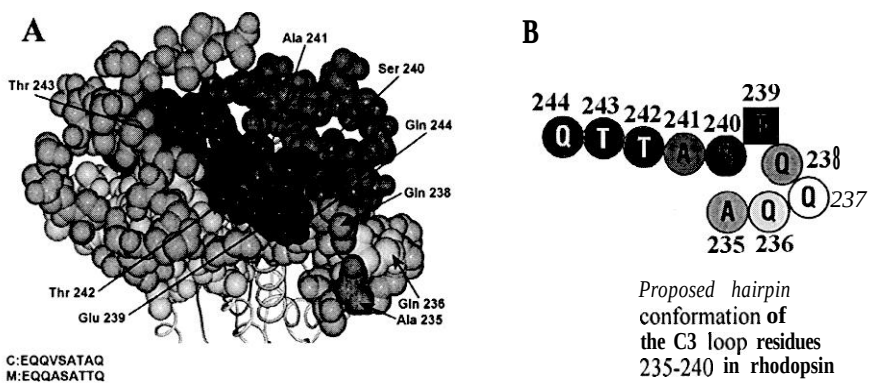


Figure 4: Mapping of the 4B4 antibody epitope for rhodopsin. Panel A shows the epitope for the antibody 4B4 mapped on the surface of a putative dark-adapted rhodopsin 3-D structure. This region is not resolved in the x-ray crystal structure, as explained in the text. The 4B4 consensus probe EQQVSATAQ was aligned (using FINDMAP) to the rhodopsin residues EQQASATTQ. Mapped epitope residues are ordered such that they follow a rainbow color scheme from red (residue one of the consensus epitope) to purple. Panel B shows the proposed reorientation of residues 235-244 of the C-3 loop of rhodopsin, based on the best-scoring FINDMAP alignment.

antibodies. The problem is to cluster these probe-target alignments into putative epitope groups and perhaps derive a confidence value for each epitope predicted.

With high throughput enhancements fully or partially in place, the antibody imprinting approach may be able to solve many otherwise intractable protein structures that are being identified in large numbers in structural genomics projects. Finally, the antibody imprinting technique can be used to assess the accuracy of computational protein structure prediction algorithms on proteins with otherwise unknown structures. Many protein structure prediction algorithms do not yield a unique prediction [2, 30] and high throughput antibody imprinting may be very well suited to eliminate incorrect computational predictions.

Acknowledgments: We are grateful to Al Jesaitis and Jim Burritt for helpful discussions and assistance with the phage display experiments.

6. REFERENCES

[1] B.W. Bailey, B.M. Mumey, PA. Hargrave, A. Arendt, O.P. Ernst, K.P. Hofmann, P.Callis, J.B. Burritt, A.J. Jesaitis, and E.A. Dratz. Structural constraints on the conformation of the cytoplasmic face of dark-adapted and light-excited rhodopsin inferred from anti-rhodopsin antibody imprints, submitted. *Protein Science, 2002.*

[2] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science,* 294(5540):93–96, 2001.

[3] C.F. Barbas, D.R. Burton, J.K. Scott, and G.K. Silverman. *Phage Display: a Laboratory Manual.* Cold Spring Harbor Laboratory Press, 2001.

[4] D. Bordo and P. Argos. Suggestions for "safe" residue substitutions in site-directed mutagenesis. *J. Mol. Biol.,* 217:721–729, 1991.

[5] C.I. Branden and *C.* Tooze. *Introduction to Protein Structure.* Garland Publishing, 1999.

[6] J.B. Burritt, C.W. Bond, K.W. Doss, and A.J. Jesaitis. Filamentous phage display of oligopeptide libraries. *Anal. Biochemistry,* 238:1–13, 1996.

[7] J.B. Burritt, S.C. Busse, D. Gizachew, E.A. Dratz, and A.J. Jesaitis. Antibody imprint of a membrane protein surface: Phagocyte flavocytochrome b. *J. Biol. Chem.,* 273:24847–24852, 1998.

[8] J. Cavanagh, A.G. Palmer III, W. Fairbrother, and N. Skelton. *Protein Nmr Spectroscopy: Principles and Practice.* Academic Press, 1996.

[9] J.M. Claverie. What if there are only 30,000 human genes? *Science,* 291:1255–1257, 2001.

[10] G.M. Clore, M.A. Robien, and A.M. Gronenborn. Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. *J. Mol. Biol.,* 231:82–102, 1993.

[11] L.L. Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.,* 285:2177–2198, 1999.

[12] T. Dandekar and I? Argos. Applying experimental data to protein fold prediction with the genetic algorithm. *Protein Engineering,* 10:877–893, 1997.

[13] C. Demangel, R.C. Maroun, S. Rouyre, C. Bon, J.C. Mazie, and V. Choumet. Combining phage display and molecular modeling to map the epitope of a neutralizing antitoxin antibody. *Eur. J. Biochem.,* 267:2345–2353, *2000.*

[14] A.M. Edwards, C.H. Arrowsmith, D. Christendat, A. Dharamsi, J.D. Friesen, J.F. Greenblatt, and M. Vedadi. Protein production: feeding the crystallographers and nmr spectroscopists. *Nat. Struct. Biol.,* pages 970-972, 2000.

[15] E. Eisenstein, G.L. Gilliland, 0. Herzberg, J. Moult, J. Orban, R.J. Poljak, L. Banerjei, D. Richardson, and A.J. Howard. Biological function made crystal clear • annotation of hypothetical proteins via structural genomics. *Current Opinions in Biotechnology,* 11(1):25–30, 2000.

[16] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the theory of NP-completeness.* W. H. Freeman and Co., 1979.

[17] T. Heiskanen, A. Lundkvist, R. Soliymani, E. Koivunen, A. Vaheri, and H. Lankinen. Phage-displayed peptides mimicking the discontinuous neutralization sites of puumala hantavirus envelope glycoproteins. *virology,* 262:321–332, 1999.

[18] H.R. Hoogenboom, A.P. deBruine, S.E. Hufton, R.M. Hoet, J.W. Arends, and R.C. Roovers. Antibody phage display technology and its applications. *Zmmunotechnology,* 4:1–20, 1998.

[19] B.A. Jameson and H. Wolf. The antigenic index: a novel algorithm for predicting antigenic determinants. *CABZOS,* 4(1):181–186, 1988.

[20] C.A. Janeway and P. Travers. *Zmmunobiology.* Current Biology Ltd., 1996.

[21] A. J. Jesaitis, D. Gizachew, E.A. Dratz, D.W. Siemsen, K.C. Stone, and J.B. Burritt. Actin surface structure revealed by antibody imprints: Evaluation of phage-display analysis of anti-actin antibodies. *Protein Science,* 8:760–770, 1999.

[22] W. Kabsch, H.G. Mannherz, D. Suck, E.F.Pai, and K.C. Holmes. Atomic structure of the actin:dnase i complex. *Nature,* 347:37–44, 1990.

[23] M.J. Marinissen and J.S. Gutkind. G-protein-coupled receptors and signaling networks: emerging paradigms. *Trends Pharmacol. Sci.,* 22(7):368–76, 2001.

[24] A. McPherson. *Crystallization of Biological Macromolecules.* Cold Springs Harbor Laboratory, 1999.

[25] H. Michel. *Crystallization of Membrane Proteins.* CRC Press, 1990.

[26] E.A. Padlan. X-ray crystallography of antibodies. *Adv. Protein Chemistry,* 49:57–133, 1996.

[27] K. Palczewski, T. Kumasaka, T. Hori, C.A. Behnke, H. Motoshima, B.A. Fox, T. Le, D.C. Teller, T. Okada, R.E. Stenkamp, M. Yamamoto, and M. Miyano. Crystal structure of rhodopsin: A g protein-coupled receptor. *Science,* 289:739–745, 2000.

[28] A. Pini, F. Viti, A. Santucci, B. Camemolla, L. Zardi, P. Neri, and D. Neri. Design and use of a phage display library. human antibodies with subnanomolar affinity against a marker of angiogenesis eluted from a two-dimensional gel. *J. Biol. Chem.,* 273:21769–21776, 1998.

[29] S.S. Sidhu, H.B. Lowman, B.C. Cunningham, and J.A. Wells. Phage display for selection of novel binding peptides. *Methods in Enzymology,* 328:333–363, *2000.*

[30] K.T. Simons, C. Strauss, and D. Baker. Prospects for ab initio protein structural genomics. *J. Mol. Biol.,* 306(5):1191–1199, 2001.

[31] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.,* 147:195–197, 1981.

[32] D.C. Teller, T. Okada, C.A. Behnke, K. Palczewski, and R.E. Stenkamp. Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of g-protein-coupled receptors (gpcrs). *Biochemistry,* 40:7761–7772, 2001.

[33] J. Wess. G-protein-coupled receptors: molecular mechanisms involved in receptor activation and selectivity of g-protein recognition. *FASEB J.,* 11(5):346–354, 1997.