A new method for mapping discontinuous antibody epitopes to reveal structural features of proteins

Brendan M. Mumey

Department of Computer Science

Montana State University

Bozeman, MT 59717-3880

Email: mumey@coe.montana.edu

Phone: 406-994-7811, Fax: 406-994-4376

Brian W. Bailey

NIH/NIAAA/DICBR/LMBB

Fluorescence Studies

Park 5 Building

12420 Parklawn Dr. MSC 8115

Bethesda, MD 20892-8115

Bonnie Kirkpatrick

Department of Computer Science

Montana State University

Bozeman, MT 59717-3880

Algirdas J. Jesaitis

Department of Microbiology

Montana State University

Bozeman, MT 59717-3520

Thomas Angel

Department of Chemistry and Biochemistry

Montana State University

Bozeman, MT 59717-3400

Edward A. Dratz

Department of Chemistry and Biochemistry

Montana State University

Bozeman, MT 59717-3400

Abstract

Antibodies that bind to protein surfaces of interest can be used to report the three dimensional structure of the protein as follows: Proteins are composed of linear polypeptide chains that fold together in complex spatial patterns to create the native protein structure. These folded structures form binding sites for antibodies. Antibody binding sites are typically "assembled" on the protein surface from segments that are far apart in the primary amino acid sequence of the target proteins. Short amino acid probe sequences that bind to the active region of each antibody, can be used as witnesses to the antibody epitope surface and these probes can be efficiently selected from random sequence peptide libraries. This paper presents a new method to align these antibody epitopes to discontinuous regions of the one-dimensional amino acid sequence of a target protein. Such alignments of the epitopes indicate how segments of the protein sequence must be folded together in space and thus provide long-range constraints for solving the 3-D protein structure. This new antibody-based approach is applicable to the large fraction of proteins that are refractory to current approaches for structure determination and has the additional advantage of requiring very small amounts of the target protein. The binding site of an antibody is a surface, not just a continuous 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. Initial results for two validation cases are presented for a graph-based protein surface neighbor mapping procedure that promises to provide additional spatial proximity information for the amino acid residues on the protein surface.

1 Antibody Epitope Mapping

Proteins are nano-machines that are constructed from long chains of amino acids (typically 100-1000 elements) using 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 [Branden and Tooze, 1999]. There are at least 30,000 different proteins in human cells [Claverie, 2001] and each protein has a folded functional structure. Whenever the 3-D folded structure of linear protein sequences can be determined this information has provided important insights into mechanisms of action and may be 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. Given sufficient protein of interest, conditions are screened to seek 3-D crystals for structure determination by x-ray diffraction, however, obtaining crystals of sufficient quality may not be possible [McPherson, 1999, Michel, 1990]. Alter-

natively, 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 [Cavanagh et al., 1996]. It is also possible to predict 3-D structures de novo from the sequence of amino acids in the protein, but the available methods for structure prediction are not very accurate unless a 3-D structure of a homologous protein is already known [Baker and Sali, 2001] (also see predictioncenter.llnl.gov).

A large fraction of protein structures of interest (50% or more) cannot be solved by the traditional approaches discussed above [Edwards et al., 2000, Eisenstein et al., 2000]. Thus, we are developing the antibody imprint method to provide structural information on difficult cases that appear refractory to traditional approaches [Burritt et al., 1998, Jesaitis et al., 1999, Bailey et al., 2003]. 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 [Burritt et al., 1998, Jesaitis et al., 1999, Demangel et al., 2000, Heiskanen et al., 1999, Bailey et al., 2003]. Antibodies tend to be highly specific for the protein structures that they recognize [Janeway and Travers, 1996]. Antibodies can either recognize continuous or discontinuous epitopes. Discontinuous epitopes provide the most useful structural information in antibody imprinting, because they can reveal distant segments of primary sequence that are in close spatial proximity on the native, folded protein. Evidence to date indicates that most antibodies recognize discontinuous epitopes on protein surfaces [Padlan, 1996]. 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 [Conte et al., 1999]. Fortunately, relatively few long-distance constraints are needed to reveal the global folding of proteins [Clore et al., 1993, Dandekar and Argos, 1997]. 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 [Bailey et al., 2003].

Briefly, the antibody imprinting method is carried out by first immobilizing antibodies (against a target of interest) on beads or in plastic wells. The immobilized antibodies are exposed to random peptide libraries so that library members which bind to the antibodies can be captured by the surface. The random peptide libraries are carried on bacteriophage (that is called "phage display" of the library), as is reviewed in the following reference [Barbas et al., 2001]. Each phage has a different peptide expressed on the surface of one of its coat proteins and there are typically $5 \cdot 10^9$ [Burritt et al., 1996] and even up to 10^{12} different peptide sequences in each library [Sidhu et al., 2000]. These probe libraries contain 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 the immobilized antibodies 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 exposed again 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 sequence that has been selected. The DNA regions of selected clones are amplified by PCR with fluorescent terminators and sequenced in a standard automated DNA sequencer. In this way, the sequence for each epitope-mimetic peptide is discovered. These individual sequences are often highly conserved and 25-100 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 an epitope-mimetic peptide can be mapped onto the sequence of the target protein in question, to recognize discontinuous epitopes that provide proximity constraints on the 3-D structure of the protein. We adopt the terminology that a peptide 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. We then describe some initial work dispensing with a consensus epitope and using individual probe-target alignments to form a surface neighbor graph, apply this approach to two validation examples, and comment on some future directions being pursued in 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 similar 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 sequence inversions) 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

[Smith and Waterman, 1981]. We have chosen an approach based on a general combinatorial alignment problem, although alternative strategies such as hidden Markov models and stochastic free grammars have been employed for related problems and could be explored.

In general, we will allow any permutation of the probe sequence to align to the underlying protein sequence[1]. Furthermore, gaps will be permitted in both probe and target sequences. Large gaps can occur when aligning the probe to the target sequence when the epitope is discontinuous. We also allow unaligned probe residues,, reflecting the possibility of a non-specific residue insertions in 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 $P_A = \{i_1 < i_2 < \cdots < i_k\}$, 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 epitope gap cost,

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

The $S(A)$ component is calculated with a substitution matrix M, similar in principle to a Dayhoff matrix, used in other protein alignment contexts. We discuss our choice of substitution matrix in the experimental results section. The substitution matrix is also used to score unaligned probe residues; if the probe residue in position i is not aligned to any target position then it is charged a penalty according to the character c occurring in position i of the probe sequence. This cost can be found in the substitution matrix, in the entry $M(c, \text{-})$. We have

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

The epitope 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 $d(x)$ 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). We have evaluated different gap cost models and discuss this point later in the results section.

[1] In some cases, e.g. membrane spanning proteins, it may be known or likely that certain regions of the target protein are inaccessible to antibodies and thus can be excluded from consideration as potential alignment positions.

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. 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 [Garey and Johnson, 1979]. 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

$$\begin{aligned} A &= U \cup \{\neg x_1, \ldots \neg x_k\} \cup \{y_1, \ldots, y_k\} \\ &\quad \cup \{\text{`}c_1\text{'}, \ldots, \text{`}c_m\text{'}\} \cup \{\text{`\#'}, \text{`}*\text{'}, \text{`@'}\}. \end{aligned}$$

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 $1 \leq i \leq n$, and $M(\cdot, \text{`}*\text{'}) = 0$ (here $\cdot$ represents any symbol). For each literal α, let $[\alpha]$ be the multiplicity of α among all clauses in C. The probe string used is

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

where

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

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 $M(\cdot, \text{-}) = -\infty$). 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.

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 [Bordo and Argos, 1991] 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. The gap penalty discussed in the text was levied against gaps in the target protein sequence that were not aligned with probe residues.

4 Experimental Results

In this section, we discuss our initial experimental results using FINDMAP and our implementation of the branch-and-bound alignment algorithm described above. 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 probability matrix to score sequence alignments. We chose the matrix shown in Figure 1, since a very similar substitution matrix was developed by Bordo and Argos [Bordo and Argos, 1991] 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. As indicated under future directions, we are in the process of obtaining an experimental substitution matrix optimized for antibody imprinting.

Recently, Jesaitis and co-workers carried out antibody imprinting using a polyclonal antibody against the ubiquitous cytoskeletal protein, actin [Jesaitis et al., 1999]. They reported the manual mapping of consensus peptides derived from phage display library selection, to complex epitopes on the surface of actin. The phage-

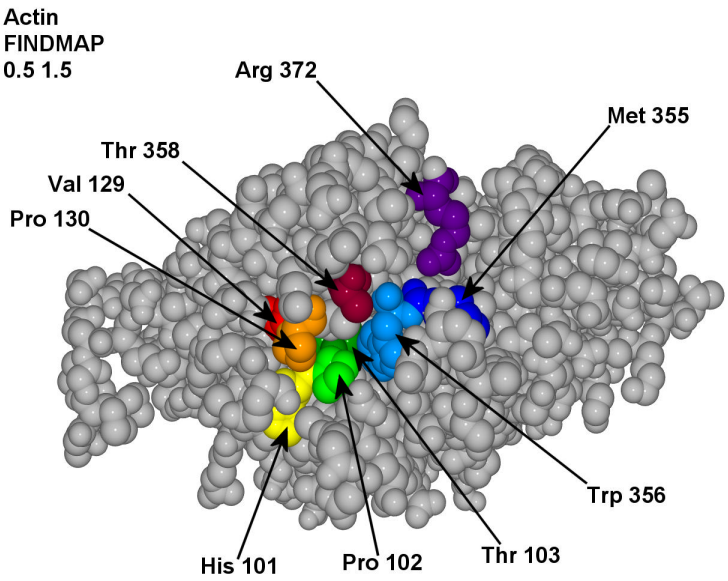
display-discovered peptides could be mapped onto the actin surface to mimic a discontinuous epitope that was consistent with the known 3-D x-ray structure of actin [Kabsch et al., 1990]. Figure 2 shows the mapping of one of the consensus sequences, VPHPTWMR, onto the surface of actin and the almost identical FINDMAP mapping. It should be emphasized that this manual mapping utilized knowledge of the actin x-ray structure and did not use residues marked with # that are not exposed on the aqueous surface in the 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 an initial validation of the antibody imprinting technique and FINDMAP on a known protein structure. Additional validation studies are described in a later section. It is possible that including an estimate for the probability of surface exposure in the overall alignment scoring function could be useful [Jameson and Wolf, 1988] and this is being explored further.

We also used the actin test case to optimize the gap cost parameters for gaps in the alignment of the target protein sequence to the probes. We chose a simple linear penalty function up to a maximum gap penalty that does not further penalize long gaps (long gaps are expected to be frequent 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, that 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, as described in the caption to Figure 2. A reason for the proliferation of near optimal solutions in this case is the freedom of the final R in the consensus probe to align to a number of positions in the target). We picked $a = 0.5$ and $b = 1.5$ from the best region for the subsequent experiments to be discussed.

The second case we considered was the integral membrane protein rhodopsin, the structure of which is not fully known. Rhodopsin is the photoreceptor for dim light vision in retinal rod cells and is an archetype for the structure and mechanism of a large superfamily of cellular G protein-coupled receptor (GPCR) proteins that re-




```
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 the anti-actin antibody epitope VPHPTWMR onto the surface of actin manually and by FINDMAP.** 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). The independent manual mapping required knowledge of the actin x-ray structure. The top-scoring FINDMAP alignment having the best match is shown above the actin sequence (#'s indicate residues known to be folded away from the aqueous surface of actin; these regions were excluded in the manual mapping).

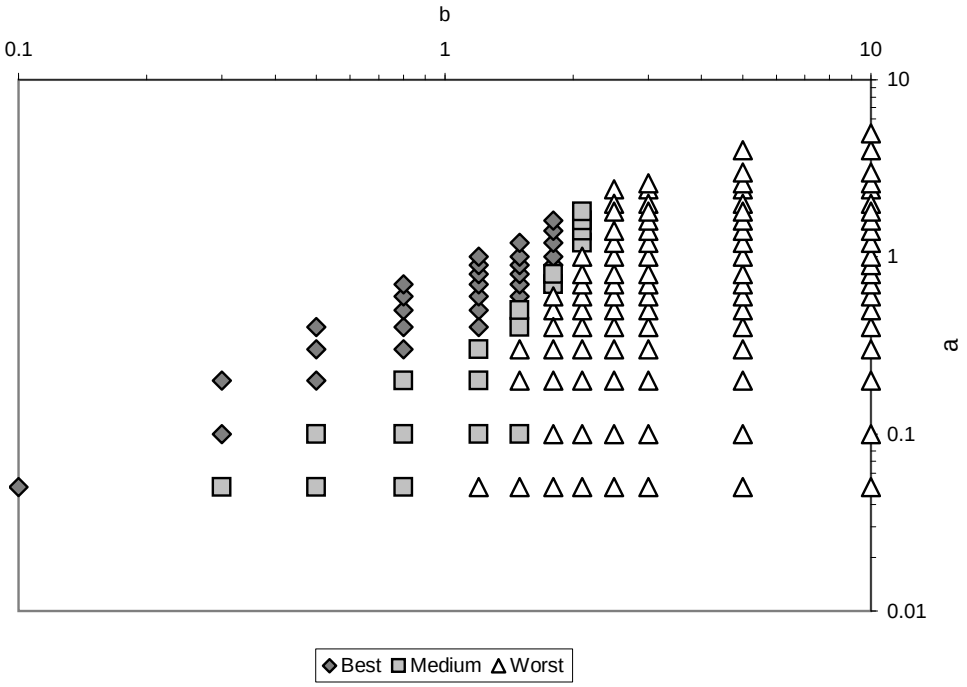


Figure 3: **FINDMAP epitope gap penalty parameter sensitivity.** For the gap distance penalty function $d(n) = \min(a \cdot n, b)$, various combinations of a and b where tested on mapping the probe epitope VPHPTWMR to the actin sequence (see Figure 2). These alignments were ranked into three categories based on how closely they agreed with the published manual mapping to the known 3-D structure of actin [Jesaitis et al., 1999]. The diamond-shaped 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, the parameter combinations used to yield the square points missed two or three residues and the triangular region more than three residues.

spond to a wide range of hormones and neurotransmitters [Wess, 1997, Marinissen and Gutkind, 2001]. The x-ray crystal structure of the dark-adapted, resting structure of rhodopsin was recently published [Palczewski et al., 2000, Teller et al., 2001] but some of the features of the protein on the cytoplasmic surface were poorly ordered in the crystals and not visible in the x-ray structure. A computational model of the missing portions of the cytoplasmic surface was built and energy minimized [Bailey, 2002]. The cytoplasmic surface structure was uncertain in the model, so antibody imprinting was applied to the aqueous surfaces [Bailey et al., 2003].

One of the antibodies investigated (B1gN) maps to the extracellular surface of rhodopsin in a compact patch that shows the proximity of two distant segments of sequence, that is in excellent agreement with a well defined region of the x-ray structure [Bailey et al., 2003] (data not shown). One of the other antibodies studied (4B4) targeted part of rhodopsin where the x-ray structure was not fully resolved, and a model of this region is shown in Figure 4A. The single most optimal mapping of the 4B4 epitope found with FINDMAP was unusual in that it was continuous with a segment of the rhodopsin sequence. The optimal mapping, however, showed a spatial discontinuity in the proximity of two parts of the epitope as illustrated in Figure 4. The aligned epitope runs from the residue colored red in Figure 4B with a rainbow color scheme, to orange, yellow and light green. The dark green residue is predicted by FINDMAP to be located spatially adjacent to the light green residue but there is a large jump in the structural model, as shown in Figure 4A. This is evidence that the surface loop folding model shown in Figure 4A is incorrect and should be adjusted to form a hairpin turn bringing A235 next to S240, as shown in Figure 4B. This example supports that notion that the antibody imprinting technique is capable of providing new structural information. Experiments are in progress to obtain more detailed conformational information by crystallizing epitope-mimetic peptides with the active site of the antibodies, to provide detailed folding information on regions of the protein surface (Lawrence, Bailey and Dratz, unpublished). The x-ray crystallography appears to be straightforward for co-crystals of peptides with antibody active sites, since molecular replacement with known antibody structures should provide the phases.

Additional antibodies will be required to reveal the complete surface structure of rhodopsin and its light-excited conformations. More detailed antibody imprinting studies, seeking to deduce light-stimulated conformational changes in rhodopsin are in progress and some of these have been submitted for publication [Bailey et al., 2003]. Most of the antibody epitopes are found to be discontinuous and thus provide important long-range distance constraints on the structures. If regions of the surfaces studied are flexible it is anticipated that a range of conformations will be deduced by different antibody epitopes consistent with that flexible structure, as would be found if other structural techniques such as NMR or x-ray diffraction could be applied. It should also be noted that it is possible to include additional information, such as from structure prediction

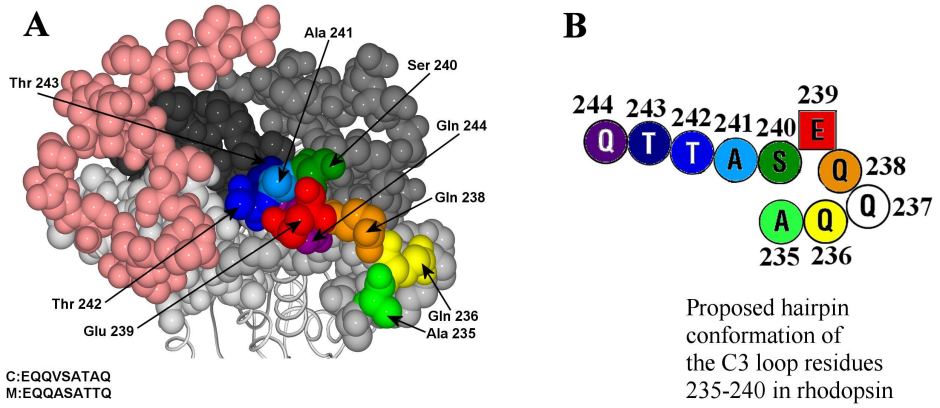


Figure 4: **Mapping of the 4B4 antibody epitope on rhodopsin.** Panel A shows the epitope of the 4B4 antibody mapped on the cytoplasmic surface of a model of the dark-adapted rhodopsin 3-D structure. This region is not resolved in the x-ray crystal structure, as explained in the text. The 4B4 consensus probe EQQVSATAQ was best aligned, using FINDMAP, to the rhodopsin residues EQQASATTQ. Mapped epitope residues are shown such that they follow a rainbow color scheme from red (the first residue of the consensus epitope) to purple. Different cytoplasmic regions of the protein are color coded: salmon residues are the C-terminal segment, dark gray is the loop between helices V and VI, medium gray is the loop between helices III and IV and light gray is the loop between helices I and II. Panel B shows the proposed reorientation of residues 235-244 of the C-3 loop of rhodopsin, with A235 moved next to S240, based on the best-scoring FINDMAP alignment, as discussed in the text.

algorithms or experimental information, if available, from intramolecular cross-links identified by mass spectrometry [Young et al., 2000] or from site-specific spin labeling [Hubbell et al., 2000] to add to the information obtained from FINDMAP or to prioritize alternative FINDMAP spatial proximity mappings.

5 A Graph-based Approach to Surface Epitope Mapping

An improvement to the overall process that we are currently pursuing is to eliminate the step of finding a consensus epitope sequence. Typically 25-100 peptide probes are sequenced that show strong affinity to the antibody in question. These sequences are often rather similar, but typically not identical. The center of the antibody combining site is expected to contribute to the highest probe affinity [Conte et al., 1999], whereas

more peripheral binding site residues tend to make a lower contribution to the affinity, and thus may show alternative binding modes. Rather than going through the step of finding a consensus sequence, FINDMAP can be run on each of the probe sequences individually to generate a family of top-scoring alignment sets, one set for each probe. These alignments are similar but often indicate the proximity of additional residues on the protein surface. We using a graph-based approach to merge and visualize the collective surface proximity information provided in the entire set of alignments. In this approach each residue of the target protein constitutes a vertex in a weighted surface-neighbor graph. Edge weights in this graph indicate how strongly the alignment data supports the conclusion that the residues at each endpoint are neighbors on the surface of the protein. The specific procedure employed for calculating edge weights is as follows: for each probe, compute the set of top-scoring alignments. Suppose there are n such alignments and that a particular pair of residues are neighbors in k of these alignments. Then k/n is added to the weight of the edge between the two residues in question. After this procedure is repeated for each probe, edges that have comparatively high weights are most likely to link residues that are true surface neighbors. In practice, errors occur both in the experimental methods used to identify the probe sequences as well as cases where the top-scoring alignments are not biologically accurate. Thus it appears useful to use a weight cutoff; edges are only kept if their weight is greater than a prespecified cutoff. If the cutoff is too low, it is likely that false surface neighbor relations will be included in the graph; too high and true neighbors will be lost. Another procedure that seems useful for pruning out non-epitope residues from the surface neighbor graph is to retain only vertices that are incident to at least one high-weighted edge. This procedure was performed on the surface neighbor graphs shown in Figure 5, only vertices incident to an edge of weight at least 50% of the maximum weight were kept. Also, a cutoff value of 1 was used to prune low weight edges.

The target sequence is also scanned for multiple occurrences of tripeptide (very rare) or dipeptide sequences in the probe and hits involving these ambiguous sequences are omitted from consideration to minimize false positive hits. It is important to eliminate false positive residue proximity information to provide accurate structure, whereas false negatives are more tolerable. An example of a surface neighbor graph based on actin FINDMAP alignment data of individual probes is shown in Figure 5A. The somewhat larger protein surface mapped with this approach, compared to Figure 2, is consistent with the fact that the antibody investigated is polyclonal. Monoclonals that we have primarily used in this work provide surface maps with a smaller area coverage, but it has been found feasible to map mixtures of several monoclonals in parallel in a single experiment (Bailey and Dratz, unpublished).

We wish to explore to what extent the surface neighbor graph can be used to make a map of the surface

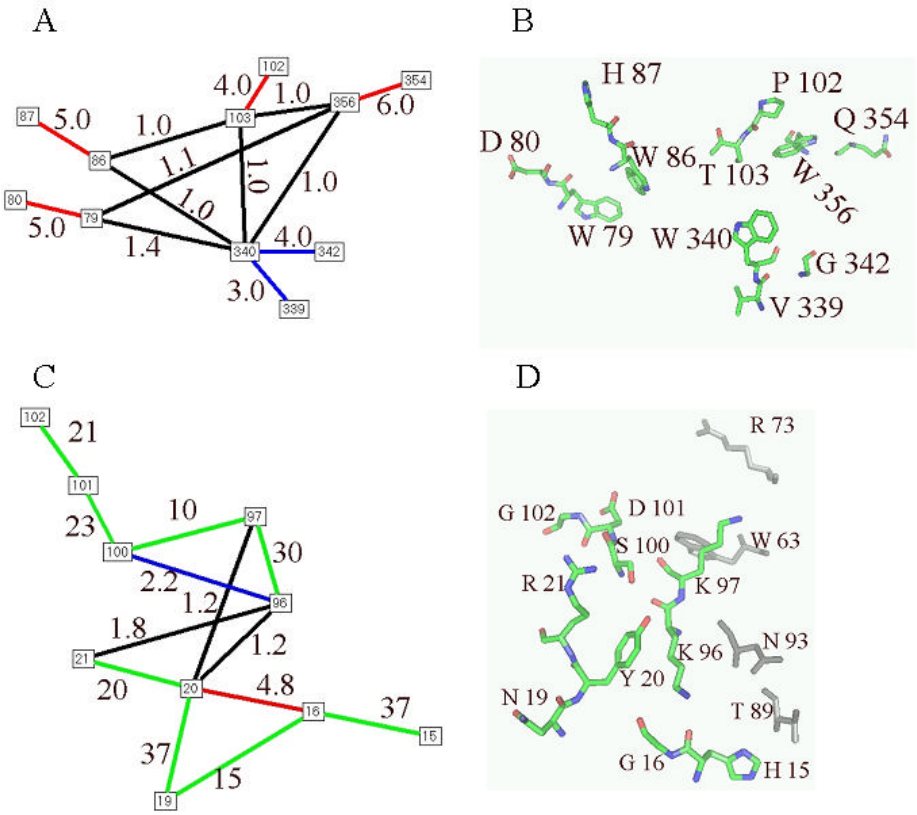


Figure 5: **Surface Neighbor Graphs and Corresponding Protein Epitopes** Panels A and C: Surface neighbor graphs contain numbers in rectangular boxes that indicate the residue sequence numbers mapped in the proteins considered. The edge weights in the graphs are color coded according to their strength. Panel A: Actin surface neighbor graph constructed as described in the text from the collection of top-scoring alignments for a set of 90 experimental probe epitope mimetic peptides found for a polyclonal antibody against actin [Jesaitis et al., 1999]. Panel B: Physical location of the residues mapped in panel A in the crystal structure of actin (PDB: 1ATN). Panel C: Surface neighbor graph of lysozyme constructed as described in the text. Note: Gap cost parameters were optimized separately for lysozyme, $a = 0.3, b = 1.0$ were found to yield the best alignments. Panel D: Known Hyhel-10 antibody epitope on lysozyme determined experimentally from the x-ray crystal structure of the complex (PDB: 1C08). The sidechain and backbone atoms of the epitope residues that were mapped in panel C are shown with nitrogen in blue, oxygen in red, and carbon in green. Residues in gray were either excluded from the mapping because of ambiguous, multiple occurrences in the target sequence (probe residues NT) or were not vertices in the surface neighbor graph because they were not connected to the main graph with sufficiently heavy edges (W63 and R73).

of the protein. The protein surface is two-dimensional, so it seems feasible to consider planar embeddings of the surface neighbor graph that place residue vertices in such a way that heavily weighted edges connect neighboring vertices in the embedding. Another criteria is that residues should be packed in a roughly uniform way, perhaps in lattices and/or proportional to their molecular volume. We are currently experimenting with various approaches to perform this embedding, including the use of some standard graph layout packages (Figure 5 was generated using the program Graphlet, available at www.fmi.uni-passau.de/Graphlet). A further important constraint is that residues that are consecutive in the linear protein sequence must also necessarily be neighbors in the embedded surface map. Other related problems that might be useful for protein surface mapping from antibody epitope data include maximum planar subgraph and minimum edge distance graph layout.

We are also 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, to more thoroughly validate the accuracy of the surface-mapping method. In these validation cases the correct antibody epitope mappings are known. The first case that we have investigated is Hen Egg Lysozyme complexed with several different monoclonal antibodies. The antibody contacts on the lysozyme surface have been identified (using the CCP4 Contacts program, http://www.ccp4.ac.uk/main.html). A collection of 50 hypothetical probe sequences were generated by randomly connecting adjacent residues on the lysozyme contact surface on the Hyhel-10 antibody. FINDMAP alignments were found to all the probes generated and the epitope surface neighbor graph was found, using the method described above. In Figure 5C we show the computed surface neighbor graph and the true epitope surface for monoclonal antibody Hyhel-10 (PDB:1C08). Also shown in Figure 5D is a diagram of the experimental monoclonal antibody epitope, that is seen to agree favorably with the surface neighbor graph edge weights.

6 Future Work

The antibody imprinting approach appears to be quite general and we are applying it to a number of additional cases. We are using antibody-protein complexes, whose 3-D structure is known, to optimize the values of the gap penalty parameters. We are further validating this approach by experimentally mapping several known anti-lysozyme epitopes and using this information to refine the substitution matrix. The validation cases are also expected to be a useful guide for more systematic choices of cutoff-weights for the planar graph models of protein surfaces as shown in Figure 5. Finally, we are applying the antibody imprinting technique to several

integral membrane proteins that are difficult structural targets [Mills et al., 1998, Burritt et al., 2001] and to reveal the nature of functional conformational changes in membrane proteins [Bailey et al., 2003].

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, the gap penalty parameters, and surface graph weights, we plan to introduce 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 protein 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 [Pini et al., 1998, Hoogenboom et al., 1998] has recently been developed and shows promise for much more rapid identification and preparation of specific antibodies. Affinity maturation steps applied to antibodies selected from random libraries to obtain subpicomolar antibody affinities [Pini et al., 1998] may also be adaptable to high throughput approaches. Random antibody libraries appear to be 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 mixtures of selected random antibody library members, avoiding the growth and screening of individual phage clones. 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 antibodies selected from a library that bind the target protein will not be known in general. Probe-target alignments can be evaluated for each probe found that binds at least one of these antibodies. The problem is to cluster these probe-target alignments into putative epitope groups and perhaps derive a confidence value for each epitope predicted. It may be possible to simultaneously collate all of probe-target alignments to produce a surface-neighbor graph that contains (possibly disconnected) clusters for each epitope present in a similar way as the example shown in Figure 5, but perhaps with a much larger surface coverage.

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. Perhaps most significantly, the antibody imprinting technique described can be used to assess the accuracy of protein structure prediction algorithms for proteins with otherwise unknown structures. Ab initio protein structure predictions are typically not unique [Baker and Sali, 2001, Simons et al., 2001]; an-

tibody imprinting promises to be an effective method to screen out incorrect predictions and arrive at more accurate folded protein structures.

Note: An initial version of this paper was presented at the RECOMB2002 conference and the expanded abstract was published in the RECOMB2002 proceedings.

Acknowledgments: The authors would like to thank Pat Callis for assistance with rhodopsin modeling and Jim Burritt for the J404 phage peptide library. This work was partially supported by grants NIH R01 GM62547 (to EAD), R01 AI22735, and R01 AI26711 (to AJJ).

References

[Bailey, 2002] Bailey, B. (2002). *Antibody Imprinting Studies of Rhodopsin, A Model of G Protein-coupled Receptor*. PhD thesis, Montana State University.

[Bailey et al., 2003] Bailey, B., Mumey, B., Hargrave, P., Arendt, A., Ernst, O., Hofmann, K., P.Callis, Burritt, J., Jesaitis, A., and Dratz, E. (2003). 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*.

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

[Barbas et al., 2001] Barbas, C., Burton, D., Scott, J., and Silverman, G. (2001). *Phage Display: a Laboratory Manual*. Cold Spring Harbor Laboratory Press.

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

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

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

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

[Burritt et al., 2001] Burritt, J., DeLeo, F., McDonald, C., Prigge, J., Dinauer, M., Nakamura, M., Nauseef, W., and Jesaitis, A. (2001). Phage display epitope mapping of human neutrophil flavocytochrome b558: Identification of two juxtaposed extracellular domains. *J.Biol.Chem.*, 276:2053–2061.

[Cavanagh et al., 1996] Cavanagh, J., III, A. P., Fairbrother, W., and Skelton, N. (1996). *Protein Nmr Spectroscopy: Principles and Practice*. Academic Press.

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

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

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

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

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

[Edwards et al., 2000] Edwards, A., Arrowsmith, C., Christendat, D., Dharamsi, A., Friesen, J., Greenblatt, J., and Vedadi, M. (2000). Protein production: feeding the crystallographers and nmr spectroscopists. *Nat. Struct. Biol.*, pages 970–972.

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

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

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

[Hoogenboom et al., 1998] Hoogenboom, H., deBruine, A., Hufton, S., Hoet, R., Arends, J., and Roovers, R. (1998). Antibody phage display technology and its applications. *Immunotechnology*, 4:1–20.

[Hubbell et al., 2000] Hubbell, W., Cafiso, D., and Altenbach, C. (2000). Identifying conformational changes with site-directed spin labeling. *Nat Struct Biol.*, 7(9):735–9.

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

[Janeway and Travers, 1996] Janeway, C. and Travers, P. (1996). *Immunobiology*. Current Biology Ltd.

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

[Kabsch et al., 1990] Kabsch, W., Mannherz, H., Suck, D., Pai, E., and Holmes, K. (1990). Atomic structure of the actin:dnase i complex. *Nature*, 347:37–44.

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

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

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

[Mills et al., 1998] Mills, J., Miettinen, H., Vlases, M., and Jesaitis, A. (1998). *Molecular and Cellular Basis of Inflammation*, chapter The structure and function of the N-formyl peptide receptor, pages 215–245. Humana Press Inc., Totowa, NJ.

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

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

[Pini et al., 1998] Pini, A., Viti, F., Santucci, A., Carnemolla, B., Zardi, L., Neri, P., and Neri, D. (1998). 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.

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

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

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

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

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

[Young et al., 2000] Young, M., Tang, N., Hempel, J., Oshiro, C., Taylor, E., Kuntz, I., Gibson, B., and Dollinger, G. (2000). High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci.*, 97(11):5802–5806.