# Computational Modeling of Peptide-Protein Binding

by

Jack Scott Emery

A Dissertation Presented in Partial Fulfilment
of the Requirements for the Degree
Doctor of Philosophy

ARIZONA STATE UNIVERSITY

December 2010

Computational Modeling of Peptide-Protein Binding

by

Jack Scott Emery

has been approved

October 2010

Graduate Supervisory Committee:

Vincent Pizziconi, Co-Chair
Neal Woodbury, Co-Chair
Eric Guilbeau
Phillip Stafford
Thomas Taylor
Bruce Towe

ACCEPTED BY THE GRADUATE COLLEGE

# ABSTRACT

Peptides offer great promise as targeted affinity ligands, but the space of possible peptide sequences is vast, making experimental identification of lead candidates expensive, difficult, and uncertain. Computational modeling can narrow the search by estimating the affinity and specificity of a given peptide in relation to a predetermined protein target. The predictive performance of computational models of interactions of intermediate-length peptides with proteins can be improved by taking into account the stochastic nature of the encounter and binding dynamics. A theoretical case is made for the hypothesis that, because of the flexibility of the peptide and the structural complexity of the target protein, interactions are best characterized by an ensemble of possible bound configurations rather than a single "lock and key" fit. A model incorporating these factors is proposed and evaluated. A comprehensive dataset of 3,924 peptide-protein interface structures was extracted from the Protein Data Bank (PDB) and descriptors were computed characterizing the geometry and energetics of each interface. The characteristics of these interfaces are shown to be generally consistent with the proposed model, and heuristics for design and selection of peptide ligands are derived. The curated and energy-minimized interface structure dataset and a relational database containing the detailed results of analysis and energy modeling are made publicly available via a web repository. A novel analytical technique based on the proposed theoretical model, Virtual Scanning Probe Mapping (VSPM), is implemented in software to analyze the interaction between a target protein of known structure and a peptide of specified sequence, producing a spatial map indicating the most likely peptide binding regions on the protein target. The resulting predictions are shown to be superior to those of two other published methods, and support the validity of the stochastic binding model.

# ACKNOWLEDGEMENTS

No acknowledgement of the many colleagues and friends who supported, encouraged, and contributed to this research can adequately capture the magnitude of the author’s debt.  Particular credit belongs to Dr. Stephen A. Johnston, who generously invested time and supplied funding to allow an inexperienced non-chemist, non-immunologist computer modeler to spend several years pursuing an unorthodox hypothesis in his Center, contributed invaluable advice and insight during many hours of discussion along the way, and, not least, invented and developed the peptide technologies that are intimately associated with this and all other work in the Center. Also in no particular order to Dr. Tom Taylor, who in addition to his many creative ideas and mathematical insights, supplied badly needed focus and discipline via more than two years of weekly meetings to review progress and plan next steps; to Dr. Vince Pizziconi, for encouraging a retired lawyer, in numerous interesting and provocative discussions, to pursue and persevere in an undertaking that has turned out to be both more demanding and more rewarding than ever imagined; to Dr. Neal Woodbury, who despite increasingly punishing management responsibilities in the Biodesign Institute, generously gave of his time to supervise initial research efforts, and provided the opportunity to become a part of the Biodesign team; to Dr. Phil Stafford, for supplying the microarray data analyzed in Chapter 2, and for many useful discussions and a great deal of patient coaching on statistical methods; to Dr. Bruce Towe and Dr. Eric Guilbeau, for their excellent instruction

and many conversations about diverse engineering topics over a period of nearly 20 years; and to all of the foregoing, for subjecting themselves to yet another dissertation committee.

To Rebecca Halperin, for the microarray equilibration data discussed in Chapter 2; to Dr. Matt Greving, for the cross-linking data discussed in Chapter 5; to Dr. Chris Diehnelt and Dr. Paul Belcher, for SPR data discussed in Chapter 2 and elsewhere. To Dr. Banu Ozkan and Dr. Steve Wells, for instruction, help, and insight on various biophysics questions.

To colleagues and friends in the Center for Innovations in Medicine, for admitting a computer geek into the company of chemists and biologists, and for the daily interactions, discussions, comments, and arguments that make up the intellectual air that we breathe, again in no particular order: Dr. Chris Diehnelt, Dr. Zhan-Gong Zhao, Dr. Kathy Sykes, Dr. Trish Carrigan, Dr. Zbig Cichacz, Dr. Chao Li, Dr. Mitch Magee, Dr. Valeriy Domenyuk, Dr. Nidhi Gupta, Dr. Bart Legutki, Dr. Lucas Jimenez, Dr. Eric Thompson, Dr. Paul Belcher, Rebecca Halperin, Preston Hunter, Kevin Brown, and Patty Madjidi.

And to my family, my son Jack and my wife Neble, who provide the motivation.

TABLE OF CONTENTS

Page

Page

Page

# LIST OF TABLES

Table Page

# LIST OF FIGURES

Figure Page

Figure Page

Figure Page

# CHAPTER 1: PEPTIDE LIGAND DISCOVERY AS AN ENGINEERING PROBLEM

A research pursuit of particular importance, and one that has attracted much commercial attention, is the effort to devise a reliable platform for constructing targeted affinity ligands having properties similar to those of naturally occurring antibodies [1-3]. Natural antibody production methods are expensive, time consuming, problematic in terms of quality assurance, not always successful, invite animal welfare-related controversy, and scale poorly [1-4]. Antibodies are large and have poor penetrative properties [5], and may themselves provoke undesirable physiological reactions [6]. The need for a practicable alternative is acute for applications such as proteomics [7], molecular imaging [1, 8], separation and purification [3, 9-11], and immunochemical assays [1], to say nothing of the boundless scope for non-immunoglobin-based ligands in diagnostics [12, 13] and therapeutics [8, 14]. In the literature can be found reports of more than fifty different scaffolds for constructing synthetic affinity ligands [1]; the extensive review by Hey, et al. [2] lists fourteen that have been launched as commercial products.

Peptides represent a class of ligands whose potential, particularly as drug candidates and in diagnostics, is only now beginning to be realized [15]. Peptides pose difficult issues in terms of delivery, stability and half-life, and bioavailability [16], but there are significant advantages to a category of compounds that offer a nearly unlimited and generally non-toxic functional repertoire [17] based entirely on relatively easily synthesized combinations of a small number of well-characterized amino acids. These properties readily lend themselves to strategies for rational design of engineered peptide-based ligands, a field that is deservedly receiving increasing effort and attention [17-27]. Nevertheless, as will be developed in the chapters that follow, peptides do not lend themselves well to the standard design strategies that have been formulated and optimized mainly in the context of small organic molecules targeted at known binding sites. In short: peptides are different.

A major research effort in the Center for Innovations in Medicine in the Biodesign Institute at Arizona State University involves an attempt to leverage the advantages of peptide ligands and multivalent binding to devise a synthetic scaffold having antibody-like affinity and specificity properties. These "synbodies" are constructed according to a template and procedure invented by the Center's Director, Dr. Stephen A. Johnston [28]. As currently conceived, construction of a synbody entails identifying two or more peptides that have intermediate affinity for the intended target, and attaching them to a linker to make a multivalent ligand suitable for use in therapeutic or diagnostic applications. [29-32]. The component peptides are chosen by screening a library (the "CIM-10K library") currently containing slightly more than 10,000 presynthesized random 20-mer peptides[1]. These are screened against targets via robotically spotted peptide microarrays or by high-throughput surface plasmon resonance (SPR) analysis. The selected peptides may then be optimized by synthesizing and testing variants with single residue substitutions, and incorporating and combining the substitutions that best improve affinity and/or specificity [33]. The selected peptides are conjugated to a linker, which determines their relative orientation and separation, and provides a scaffold that can be functionalized to facilitate detection, surface attachment, or delivery of a therapeutic payload. The intent is for the component peptides to bind the targeted protein simultaneously at distinct sites, providing an antibody-like affinity commensurate with the sum of the binding energies of the individual peptides. Specificity should also be enhanced due to the improbability of both peptides simultaneously having high affinities for targets other than the one against which they were selected.

Herein lies an engineering challenge. Engineering design implies calculation; we do not build jetliners by bolting random parts together. Can the process of identifying high affinity peptides be made more a matter of engineering design based on quantifiable properties, and less a matter of discovery by trial and

[1] The three C-terminal residues are always GSC or GSG, depending on the application. The other 17 positions are determined computationally by a pseudorandom process with residues selected at equal probability from the set of 20 naturally occurring amino acids excluding cysteine (to avoid disulfide bridges). Where present, the C-terminal cysteine is used for attachment to the array surface.

error? The goal of the research presented here is to begin assembling the tools and theoretical insights needed to evaluate in advance, through computational modeling of the physics and geometry of the interaction between a candidate peptide and protein target of known structure, the likelihood that a given peptide will display the affinity, specificity, conformational geometry, and cooperative binding properties needed for the intended application.

An ideal end point would be an algorithm capable of specifying, by pure computation, the composition of peptides and linkers meeting stated requirements, leaving nothing for the chemists to do but assemble the indicated parts.  Although the job security of the chemists is not yet in jeopardy, theoretical and computational approaches can nevertheless make such potentially useful contributions as reducing the size of the library that must be screened; suggesting ways to optimize library composition; providing improved criteria for ranking the peptides identified by screening; and elucidating peptides' binding position and geometry so as to better inform the selection of suitable linkers.

Under the synbody design process currently in use, the affinity and specificity of peptides for the intended target are inferred from microarray or SPR assays, subject to considerable experimental uncertainty and at a cost of screening a large number of candidate peptides against multiple targets and/or in the presence of competitors. Given the large number of possible peptide sequences – about 5.5 x $10^{21}$ for 20-mer peptides with 17 variable positions and a 19 amino acid alphabet – it is impracticable to search or sample more than an infinitesimal fraction of the sequence space. Taking into account the high cost of synthesizing and testing candidate binding elements, there is value in any strategy that can focus the screening process on the better candidates and avoid peptides whose unsuitability can be predicted *a priori*.

Cooperativity -- the ability of the selected peptides, once conjugated to the linker, to bind the target simultaneously without undue loss of affinity -- is a requirement for which no means of evaluation exists in the current process, short of actually constructing and testing synbodies corresponding to each possible combination of peptides and linker.  The goal of multivalent binding implies a need to select peptides capable of binding at distinct loci, and in positions such that the spacing and orientation of the bound peptides is compatible with the dimensions and geometry of the linker. Pairs of peptides that compete for the same binding site, or that prefer to bind in positions or orientations that the linker cannot comfortably span, are unlikely to bind cooperatively when linked. Therefore, a particular focus of the work described here (see Chapters 4 and 5) has been the computational prediction of peptide binding loci on protein targets for which a solved structure is available.

The task of devising predictive models of peptide-protein binding is made more challenging by the considerable uncertainty surrounding a number of theoretical questions that relate mainly to the high flexibility of peptides in the size range of interest, which arises from the many rotatable bonds present.  How flexible is a typical 20-mer peptide? To what extent do these peptides assume stable conformations in solution? Can binding occur only if the peptide is in a conformation that "fits" a specific binding site on the target protein, or can the peptide bind partially or suboptimally and then adjust its conformation and position? Do peptides typically bind in a single fixed position, or can they move around on the binding site? Is peptide binding best characterized in terms of a single binding site on the target protein, or are there multiple sites to any of which alternate conformations of the peptide may bind?

Obviously, the mechanism and kinetics of binding, and therefore the affinity and specificity, depend to a considerable extent on the answers to questions of this kind. For example, the entropic penalty on binding may be very large if peptides are completely flexible in solution and held rigidly when bound, or it may be modest if peptides assume stable conformations in solution and are relatively mobile when bound. And assessing the aggregate contribution to binding energy of all of the non-bonded interactions – hydrogen bonds, salt bridges, hydrophobic forces, etc. -- occurring between the bound peptide and the protein becomes much more challenging if the peptide is dynamically making and breaking interactions as it changes position.

Chapter 2 addresses these and other matters of theoretical background, presents model results relating to peptide flexibility and binding kinetics, and concludes by proposing a hypothetical mechanism of peptide binding. In the proposed model, peptide binding is characterized not in terms of the usual shape and charge-dependent "lock and key" metaphor, but as a rather mobile and dynamic interaction, best thought of in more probabilistic terms, in which the peptide may explore distributions of positions and conformations around multiple energy minima.

Chapter 3 summarizes the results of a comprehensive analysis of the properties of the peptide-protein interfaces for which structures are available in the Protein Data Bank (PDB), from which are derived heuristics relating to the characteristics of peptides most relevant to their protein-binding behavior. These will be seen to be generally consistent with the proposed theoretical model. The results of this analysis have been distilled in a database, made publicly available, that includes detailed descriptors of geometry, energetics, and non-bonded interactions, for 3,924 peptide-protein interfaces extracted and curated from the PDB. In Chapter 4, a novel computational technique will be presented for evaluating and spatially mapping the chemical and/or interactive properties of protein surfaces, and applied to the problem of predicting peptide binding loci; in Chapter 5, the predictions of this method will be compared with an experimental determination (by others) of binding loci of synbody peptides on AKT-1 protein. Chapter 6 will draw conclusions regarding the strengths and weaknesses of the model developed in Chapters 2 through 5, and suggest improvements.

## CHAPTER 2: THEORETICAL BACKGROUND AND LITERATURE REVIEW

It appears that no general theory of peptide-protein binding yet exists. An abundance of information can be found pertaining to particular classes of interactions involving specific protein types and known binding sites [25, 34-48]. A very large body of literature exists, for example, on prediction of peptide binding to MHC complexes [35, 36, 38-40, 44, 46, 48], using a variety of approaches, including bioinformatic or statistically based prediction [35, 37], QSAR [39], machine learning [47], structural complementarity-based prediction [25, 38], molecular dynamics simulation[38], and energy optimization [44]. Similar techniques have been used to predict binding of peptides to G protein receptors [41], calmodulin [42], SHC and PDZ domains [47], and other proteins having known peptide binding domains. These studies are at best tangential to the problem addressed here, since the proteins of interest as synbody targets would rarely have even one known peptide binding site, much less two. It is therefore necessary to attempt to construct a predictive strategy from a point of departure closer to first principles.

The canonical expression for the affinity of a ligand for a target is the familiar relation $\Delta G = RT \log (K_D)$, where $K_D$ is the dissociation constant, T is absolute temperature, R is the gas constant, and $\Delta G$ is the change in Gibbs free energy [49-51]. Therefore, one way of predicting affinity is to estimate the change in Gibbs free energy based on a suitable computational model. Formally,

$\Delta G$ comprises an enthalpic and an entropic term: $\Delta G = \Delta H - T\Delta S$, where $\Delta H$ is the change in enthalpy, T is the absolute temperature, and $\Delta S$ is the change in entropy between the two states for which $\Delta G$ is to be determined. For modeling purposes, the Gibbs free energy can be decomposed in terms of the energy contributions of the various forces and interactions present [52], and/or as a function of selected descriptors whose contributions or weights are determined by fitting to a training set [53-55]. At 30°C, an affinity of 10 μM ($K_D$) corresponds to $\Delta G$ of approximately -7 kcal/mol, which can be accounted for, at least in theory, by a very few hydrogen bonds (each ~2 to 7 kcal/mole), salt bridges (~3 to 5 kcal/mole), and/or hydrophobic interactions (~0.5 to 3 kcal/mole).

As a foundation for a predictive model of peptide binding, the relation between $\Delta G$ and $K_D$ presents several challenges. As will be shown below, a peptide-protein interaction arguably cannot be represented accurately by a single peptide conformation in a single bound position, but instead requires consideration of an ensemble of states representing a distribution of possible bound configurations. The possibility of transitions between these states introduces the further potential complication of multi-step kinetics. Moreover, there are a great many factors that potentially influence $\Delta G$. One recent study tested more than 60 different descriptors in an effort to discover an optimal set for use in a parameterized binding energy model [54]. Many of these factors depend very sensitively on the exact atomic-level dimensions and geometry of the interface, which are at best measurable only approximately by standard x-ray crystallography and nuclear magnetic resonance (NMR) techniques.

These and related theoretical challenges are the primary focus of the sections to follow. The first section will discuss experimental results and theoretical considerations indicating the extent of conformational flexibility of peptides in general and the intermediate-sized random-sequence peptides of the kind employed for synbodies in particular, focusing mainly on the unbound state. The implications of this flexibility with regard to the mechanism and kinetics of peptide binding to proteins are then addressed. The chapter will conclude with a discussion of what these properties imply regarding the construction of a predictive model of peptide binding.

### The nature and extent of peptide flexibility

The discussion to follow will begin with a brief review of the basic geometric and physical properties that determine peptide structure, followed by a summary of the literature on peptide structure prediction. Experimental evidence will be presented suggesting that, at least with respect to the intermediate-sized random peptides here of interest, considerable peptide flexibility must be assumed in order to account for the binding observed in library screening experiments. Theoretical evidence and molecular dynamics modeling results will then be presented confirming that the peptides of interest appear to occupy a distribution of conformations in the unbound solvated state.

### The structural characteristics of peptides generally

Peptides are polymers consisting of amino acids linked by peptide bonds. Each of the 20 naturally occurring amino acid species has a characteristic side chain extending from the $C_\alpha$ atom. The overall structural shape assumed by the peptide molecule is determined by a linear "backbone" consisting of the C, N and $C_\alpha$ atoms and the bonds between them, with side chains extending as branches from the $C_\alpha$ atom at each residue position (except glycine, which has no side chain, and proline, whose side chain forms a re-entrant ring rather than a branch).

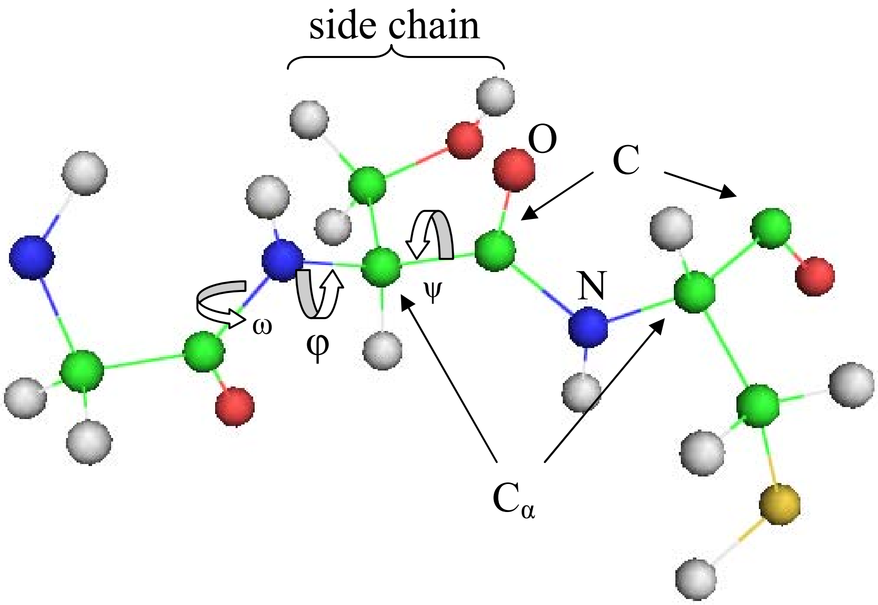

Figure 1. Structural features of a trimer peptide segment.

The geometry of the peptide backbone is determined by the three dihedral angles φ, ψ, and ω; the interatomic distances between N and $C_\alpha$ and between $C_\alpha$

and C, and the peptide bond length; and the three bond angles. See Figure 1. The interatomic distances and bond angles vary only slightly, and the peptide bond between C and N is relatively planar with the carbonyl oxygen trans to $C_\alpha$ in more than 99% of cases, so variation in $\varphi$ and $\psi$ accounts for most of the backbone flexibility of peptide structures [56]. The $\varphi$ and $\psi$ torsions are not constrained by the electronic structure of the bonds, but, as will be seen, rotations about the N - $C_\alpha$ ($\varphi$) and $C_\alpha$ - C ($\psi$) bonds are significantly limited by steric constraints and by the tendency to prefer orientations that occupy rotational energy minima. Proline residues impose a relatively rigid backbone bend, since the cyclic side chain constrains rotation about the N - $C_\alpha$ bond.

The degree of conformational diversity available to side chains depends on the residue type. Again, bond lengths and bond angles may be regarded as essentially fixed. Some residues (glycine, proline, and, neglecting hydrogen positions, alanine) have no rotatable bonds (although the proline ring can adopt either of two main conformations), and others (lysine and arginine) have as many as five (again neglecting rotations affecting only hydrogen positions). Rotamer statistics determined by molecular dynamics simulation and/or compiled from known protein structures indicate that a limited set of preferred side chain conformations predominates [57, 58].

A further potential determinant of polypeptide structure, highly significant in folded proteins, is the tendency for non-bonded interactions such as hydrogen bonds to occur between disparate parts of the chain, tending to stabilize folded conformations that foster formation of such contacts. Polypeptide chains also tend to favor conformations that hide hydrophobic moieties from the solvent (here assumed aqueous). A crucial issue in peptide modeling is the extent to which the conformation of a given peptide is likely to be influenced by these factors. It is well established (and obvious from the geometry) that the contact density (number of intra-chain contacts per residue), and therefore the opportunity for intra-chain non-bonded interactions, increases quite rapidly with chain size for chain lengths up to approximately 100 [59]. The conformational entropy, meanwhile, increases approximately linearly with the number of rotatable bonds. Therefore, as chain lengths become shorter, it becomes much less likely that folded conformations can be found in which the improvement in ΔH from intra-chain interactions can overcome the entropic penalty due to the restriction of rotational degrees of freedom. For peptides whose sequences do not lend themselves to sufficient high-enthalpy contacts, as for intrinsically unstructured proteins, the TΔS component of free energy is likely to assume relatively greater importance than is the case for folded proteins, in which case the structure corresponding to the global thermal energy minimum does not necessarily represent even the predominant conformation, much less the only one [60].

Because peptide backbone bond angles and bond lengths are essentially fixed, the distance between the $C_\alpha$ atoms of adjacent residues is an approximately constant ~3.8Ǻ. A fully extended 20 residue chain is therefore approximately 7.2 nm in length, as measured between the terminal $C_\alpha$ atoms. The average end-to-end length of a polymer chain in solution depends upon the temperature and solvent conditions ([51] at 621-26; [61]). In “good solvent” conditions, interactions between monomers and the solvent are more favorable on average than interactions between monomers, and an expanded conformation results. In a “poor solvent”, monomers prefer to interact with each other, and the chain tends to collapse. The state in which these tendencies are in balance is referred to as the “theta” state; in this state, conformational entropy is maximized and the polymer chain is in a “random flight” configuration in which the average end-to-end length is the length of individual subunits (here 3.8Ǻ) times $N^{1/2}$, the square root of the number of subunits [51, 62]. For given solvent conditions, there is a temperature θ corresponding to the theta state, above which the polymer chain tends to extend, and below which it tends to compact. The sharpness of the compact-to-extended transition, in terms of the temperature range over which it occurs, depends mainly on chain flexibility [63] and chain length. For long chains (i.e. proteins), the transition may be very sharp, occurring over less than 1°C; no data was found specifically for intermediate length peptides, but the theoretical transition for polymers of similar molecular weight is much broader (at least tens of degrees) [64].

The outcome of the trade-off between entropy loss and enthalpy gain is obviously not determined by chain length and temperature alone, and theoretical polymer behavior provides at most an indication of general tendencies. The forces that determine whether a given polypeptide chain folds into a stable structure are complex, and depend sensitively on the composition and geometry of the particular sequence. Even some quite short peptides clearly do fold into stable structures [65-67], and others appear at least to display strong secondary structure tendencies [68]. The conformations observed in x-ray structures of peptide-protein interfaces are often taken as further evidence that intermediate length peptides adopt stable structures, at least upon binding to a protein [69]. It will be shown in Chapter 3 that the x-ray data actually evidences greater conformational diversity than usually supposed, a conclusion also supported by molecular dynamics modeling of peptide-MHC-I complexes [38]. (Conclusions drawn from x-ray structure data are also arguably flawed by selection bias since only ordered structures are typically solvable.) On the other hand, a rather large fraction of known protein sequence space involves segments that are intrinsically unstructured [70-72]; by some estimates, about 50 percent of mammalian proteins have disordered regions at least 30 residues in length [73]. So it is by no means a foregone conclusion that intermediate-length peptides would or would not necessarily fold, although it is now generally recognized that "in contrast to proteins, short peptides do not systematically adopt stable well-defined tertiary structures" [74]. It also seems reasonable to hypothesize that intermediate-length peptides whose sequences are of biological origin, such as those derived from subsequences of known proteins, may be more likely to fold into stable structures than the random sequence peptides of interest here, since the latter have not evolved under selection for functions that depend on structure.

For purposes of designing peptide ligands suitable for incorporation into synbodies, arguably the objective should be to select peptides that occupy the middle ground between folded and unstructured. These would present a reasonable degree of conformational diversity in solution so as to increase the probability of a productive encounter with the target, while minimizing the entropic penalty on binding as much as possible. These considerations will be developed in greater detail in the section on binding kinetics below.

### Computational prediction of peptide structure

Much progress has been made in computational protein structure prediction, as described in recent reviews [75, 76]; the most important prediction strategies are ably summarized by Nicosia, et al. [77]. The pace of advance is evident from the results of the biennial Critical Assessment of Protein Structure Prediction (CASP) competitions [78], in which research groups submit predictions after being given the sequences of proteins whose structures have been solved but not yet published. The current version of the perennial CASP winner, Rosetta, created by the David Baker group at University of Washington, routinely predicts backbone structure to within 5Ǻ root mean square deviation (RMSD) for protein domains up to 125 residues in length [79]. Another approach, comparative modeling, which makes predictions on the basis of the known structures of homologous sequences, can achieve accuracies as high as 1-2Ǻ RMSD in cases where highly homologous structures are available [75].

If the “thermodynamic hypothesis” -- that the native folded state of a protein is the lowest potential energy state [77] -- is accepted, then protein folding is, in concept, a simple problem, given an accurate energy function and infinite computing power: simply search through all possible conformations for the global energy minimum. Unfortunately, energy functions are not perfectly accurate, and improvements in accuracy (e.g. explicit solvent models) come at a cost that may easily amount to several orders of magnitude of increase in computation time required. And even if an energy function existed that could compute accurate energies instantaneously, the conformation space of a typical protein is far too large to be exhaustively searched. Therefore, success in protein folding requires finding strategies for confining the search to smaller regions of conformation space, and/or directing the search along trajectories most likely to lead to or near a global energy minimum. A few purists attack the problem with entirely physics-based algorithms; here success depends on improved understanding of folding mechanisms, so as to guide the search along a folding trajectory hopefully corresponding closely to that followed by the actual protein as it folds [80-82]. Most of the more successful protein structure prediction algorithms, including Rosetta, instead seek to confine the search by focusing first on fragments of manageable size [83-88], whose structural preferences can be sampled from known structures and/or estimated by molecular dynamics simulation [89, 90] or Monte Carlo search methods [60], perhaps using simplified energy functions that substitute aggregate descriptors for some of the atomic-level detail [91, 92].

Computational estimation of the secondary structure tendencies of fragments, for which a variety of techniques have been suggested [93, 94], and homology modeling, in which the structures of homologous sequences are used as a source of guidance [75, 95, 96], may provide additional inputs useful for directing the search to the highest likelihood regions of conformation space [90, 97].

Although computational strategies for predicting peptide structure from sequence have tended to borrow heavily from these standard protein structure prediction concepts, peptides differ from proteins in ways that present both opportunities and challenges. One obvious opportunity is that because peptides have very many fewer rotational degrees of freedom than full-length proteins, computationally intensive methods such as molecular dynamics modeling [98-101], Monte Carlo search [69, 102, 103], tabu search [104], and simulated annealing [105], become somewhat more practicable. The principal challenge, one that makes solutions obtained from the standard protein folding algorithms suspect, is the likelihood that a peptide may occupy a distribution of conformations in which intra-chain interactions are transient and relatively unimportant, rather than a single folded shape held in place mainly by stable non-bonded contacts.

A variety of algorithms for predicting peptide structures and/or ensembles of preferred conformations have been tried, with varying degrees of success (e.g. [74, 77, 105-113]). Until relatively recently, methods were often used that, as with protein folding, implicitly or explicitly conceptualized the problem in terms of a single structure corresponding to a single global energy minimum [85, 114-117]. For example, high on any list of tools for polypeptide structural analysis is Robetta, a publicly available web server implementation of the very successful and well-tested Rosetta algorithm, which, though designed for protein folding, will accept a peptide sequence as input and produce a selection of candidate folded structures [85, 103, 114, 115]. As already noted, however, methods that treat peptides as "little proteins" are likely to produce misleading results when applied to sequences that remain wholly or partially disordered in solution, although they may be useful for analyzing peptides that adopt stable folded structures. In actual comparisons with peptide NMR structures, Robetta tended to produce overly compacted conformations bearing little resemblance to the NMR models (see [69], Fig. 2).

The current state of the art of peptide structure prediction appears to be defined by three software tools specifically designed for the task -- PepStr [116, 117], PepLook [69, 118], and PepFold [69, 118] – and by a handful of other computational approaches, typically leveraging one or more existing molecular analysis packages in a one-off analysis.

The oldest of the server-based tools, PepStr [116, 117], is aimed primarily at predicting the tertiary structures of bioactive peptides, whose activity often seems to imply a single predominant structure. Pepstr generates a starting backbone conformation by predicting secondary structure and β-turn characteristics, assigns side-chain angles from a rotamer library, and uses molecular dynamics / energy minimization to arrive at a single predicted structure.

PepLook is a commercial peptide structure prediction tool that identifies a “prime” structure and computes an index of its structural stability by generating and comparing approximately 100 other low energy structures [69, 118]. At each iteration, it generates a large number of structures using φ/ψ pairs randomly chosen from a list of 64 preferred combinations. After energy evaluation, the selection of φ/ψ pairs for the next iteration is optimized by reweighting the probability of selection for each pair according to the relative proportion of lowest energy structures in which it appeared. After 100 to 500 iterations, the best 99 models are selected, energy minimized, and reported. The PepLook software is not available to end users; instead, the sequence and conditions of interest must be submitted to the company (Biosiris), and a quotation requested. Although the computational strategy appears reasonable, no reported analyses using PepLook were found other than in the two papers by its authors. In these, results are described for only five peptides, and RMSD comparisons with NMR structures are given for only three of those, claiming RMSD of 1.3Ǻ, 2.5Ǻ, and 0.8Ǻ for three peptides of length 23, 27, and 20 residues respectively. These results are difficult to interpret since it is unclear which of the 99 output models were compared with which of the (typically many) models present in the NMR structures.

The most recent entrant, PepFold [74, 113], begins by extending a concept that has found wide usage in protein structure prediction: the structural alphabet [119-123]. An example of a very simple structural alphabet is the three-letter encoding in which each residue of a sequence is assigned one of the letters 'a', 'b', or 'l' (corresponding to α-helical, β-strand, or left-handed helical, respectively) based on its $\varphi$ and $\psi$ angles [68, 124]. The sequence obtained by assigning one of these structural alphabet letters to each residue provides a rough descriptor of structure. In an analogous fashion, it is possible to construct a larger alphabet in which the assignment of letters is based on the geometry of several adjacent residues taken as a unit. PepFold uses a 27-letter alphabet to describe the geometry of overlapping 4-residue segments in terms of the six $\varphi$ and $\psi$ angles internal to the segment; in effect, the letters partition the six-dimensional space $\{\psi_1, \varphi_2, \psi_2, \varphi_3, \psi_3, \varphi_4\}$ into 27 partitions, each of which is assigned a letter. Determining the optimally informative partition boundaries is obviously a significant challenge; the PepFold authors say only that their structural alphabet was derived from a hidden Markov model.

PepFold analyzes a peptide using a classifier trained on a large training set derived from PDB structures of proteins. The classifier takes as input an 8-residue amino acid sequence (the 4-residue sequence to be assigned a structural alphabet letter plus the two flanking residues on either side), and outputs the probabilities that the 4-residue sequence would match the structural criteria of each of the 27 possible structural alphabet letters. PepFold retains the most probable letters for each overlapping 4-mer in the sequence whose structure is to be predicted, and, starting at a random position in the sequence, iteratively adds letters in both directions until arriving at the most probable structural alphabet sequences. From these, conformations are generated and the most energetically favorable are identified via Monte Carlo search using a standard force field for energy evaluation. PepFold's output is a set of one or more clusters of conformations.

PepFold appears to be the best of the current freely available analysis tools. For each of the peptides in the PepStr test set [117] of 42 bioactive peptides -- intermediate-size peptides (length 9 to 20 residues) whose structural tendencies are reasonably well represented by folded protein structures -- the PepFold solution corresponding to the most populated output cluster of conformations is close to the reference NMR model (average 2.8Ǻ $C_\alpha$ RMSD) [113]. By way of comparison, PepStr achieved an average 4.0Ǻ $C_\alpha$ RMSD on the same test set [77].

It may be doubted, however, whether PepFold's predictions would provide an accurate depiction of the conformational ensembles inhabited by the random sequence peptides of interest here. PepFold was trained on folded protein structures, and tested against NMR structures using two test sets. The first contained 10 short peptides (length 10 to 23) and 13 "mini-proteins" (length 27 to 49); the second was the PepStr test set already described. Both test sets represent biologically relevant molecules having sufficient structure to possess at least a relatively invariant rigid core.

Several research groups have published peptide structure prediction results obtained by chaining together pre-existing software tools to perform some or most of the analysis steps. Klepeis, et al. [125], pioneered the approach of performing a global optimization search on a search space defined by a standard energy function.  Applying a branch-and-bound search algorithm to search for a global energy minimum on an objective function adapted from the widely cited function Empirical Conformation Energy Program for Peptides (ECEPP) [126], they were able to obtain solution conformations in excellent agreement with experimental results for two 5-mer peptides ($C_\alpha$ RMSD < 1.5Ǻ) and one alpha-helical 10-mer peptide, decaglycine ($C_\alpha$ RMSD of 0.136Ǻ). All of the peptides studied are believed to be highly structured in solution, and only the (assumed) single native conformation was sought or reported.

The recent analysis by Nicosia, et al. [77] appears to represent the high water mark of the global energy optimization approach. Using a "generalized pattern search" algorithm (a complex procedure for discretizing the search domain and combining coarse grained global search with finer grained local search, explained in detail in [127]) Nicosia, et al. achieved ~20% improvement in predictive performance over the PepStr algorithm on the PepStr dataset.  The approach is noteworthy for its ability to proceed without the benefit of any statistical or bioinformatic inputs (although secondary structure, β-turn, and other similar constraints can be incorporated where available).  The algorithm is used to search for (in this case) energy minima in a space defined by ECEPP v. 3 [126], which computes potential energy by summing electrostatic, Lennard-Jones, hydrogen bond, and torsion energy terms (here the ECEPP function is supplemented by adding a hydration energy term). The search procedure generates many trial conformations as it iterates, and these can be clustered and analyzed. Although Nicosia, et al., focused on determining a single preferred conformation for each peptide, they note the possibility of "return[ing] the representative conformation of each cluster rather than just the conformation with the lowest potential energy value." [77]

Molecular dynamics provides another possible direction from which to approach the prediction problem. Ideally, one could begin with an extended conformation and allow the peptide to follow its natural trajectory, which should reach a steady state corresponding to the native folded conformation if one exists [68, 128]. If the peptide remains disordered, conformations can be sampled over the course of a suitably long trajectory and clustered. The molecular dynamics approach will be discussed in greater detail in the next section, and clustering data obtained from molecular dynamics experiments on a synbody peptide will be presented.

Taken together, the reported peptide structure prediction efforts of various groups just described arguably provide inferential support for the hypothesis that intermediate-length peptides in aqueous solution are likely to exist in a range of conformations rather than in a single folded state. In the literature can be seen a kind of paradigmatic evolution, beginning with an often explicit assumption of a single native structure in earlier publications, and leading to an acknowledgement in the most recent papers that many peptides are likely disordered. The two approaches (PepLook and PepFold) claiming the best predictive results both find it necessary to output multiple solution conformations, and single-solution approaches appear to make significantly poorer predictions, at least in the relatively few cases for which such comparisons have been found [69, 113]. This is so even though predictive accuracy is typically determined in comparison to NMR models, nearly all of which involve molecules that show considerable structure. At least in the case of PepFold, the typical computed solution conformations are essentially indistinguishable in terms of predicted energies if error ranges are taken into account (differences on the order of 0.5 kcal/mol or less). For their part, the PepLook authors affirmatively acknowledge that "in many cases . . . structures obtained by NMR and from in silico calculations were eventually divergent," and explain this by noting that "divergent data may be confusing if one is willing to believe that a peptide structure must be unique . . . but no longer confusing if we accept structural diversity or, in other words peptide polymorphism or disorder." ([69] at p. 895).

### Peptide conformational freedom in the unbound state

The $\varphi$ and $\psi$ backbone dihedral angles of a polypeptide chain refer to single bonds between the tetrahedral $sp^3$-hybridized $C_\alpha$ atom and the planar $sp^2$-hybridized N or C atom, respectively [129]. Although the bonds themselves are fully free to rotate, two factors significantly restrict that freedom. First, many otherwise possible rotations are disallowed because they would cause steric collisions. According to one source, "three quarters of the possible (φ, ψ) combinations are excluded simply by local steric clashes." [130] Second, because of repulsive forces between "1-4 pairs" of atoms – pairings between the atoms attached to two opposing sigma-bonded atoms – the rotational energy is higher in rotations that force these atoms closer together, so rotations that tend to place 1-4 pairs in trans are energetically favored, although the energy barriers are less than those seen between the trans / gauche$^+$ / gauche$^-$ energy minima seen for rotations about bonds between two sp$^3$-hybridized atoms [129]. These steric restrictions and rotational energy barriers partially account for the limited range of preferred dihedral angles seen in the familiar Ramachandran plots of φ and ψ frequencies in proteins (also influenced by the effects of non-bonded interactions, particularly hydrogen bonds, on folding preferences).

Approximate bounds can be placed on the extent of backbone diversity by estimating the number of meaningfully distinct rotation positions possible at each φ and ψ dihedral. Structural alphabets, described in the preceding section, provide a starting point. A rough lower bound on the number of distinct conformations corresponding to locally permissible φ and ψ rotations might be taken as the number of possible distinct combinations of the three-letter a-b-l alphabet, in which each internal residue is assigned to the right-handed helical (a), beta strand (b), or left-handed helical (l) partition of Ramachandran space [68]. In that case, for a 20-mer peptide not containing proline, the number of possible distinct backbone conformations would be $3^{18}$, or $3.87 \times 10^{8}$. More sophisticated analyses reach a finer partitioning on a conformation space of larger (typically 4- to 6-mer) fragments so as to optimize the representation of the conformational attractors obtained from statistical sampling from known protein structures. The recent study by Pandini, et al. [124] is typical of this approach; a 25-letter structural alphabet was found to optimally capture the conformational tendencies of 4-residue fragments. On that basis, a 20-mer peptide might be approximately represented by six 4-residue fragments, and the number of distinct conformations would be $25^{6}$, or $2.4 \times 10^{8}$. For an upper bound, one might assume that each of the $\varphi$ and $\psi$ dihedrals could take one of three possible positions, corresponding to assumed rotational energy minima at which the atoms attached to the C or N atom are at greatest distance from the atoms attached to the $C_{\alpha}$ atom; then the number of possible conformations for a 20-mer would be $3^{38}$, or $\sim 1.3 \times 10^{18}$.

The foregoing estimates suffer from several obvious inaccuracies. Since the structural alphabet categories are derived by sampling from protein structures, they implicitly incorporate folding-related structural constraints not necessarily present in short peptides. More importantly, it is likely that estimates of this kind very greatly overestimate conformational diversity because they do not take non-local steric collisions into account. It is difficult to quantify the latter effect, but some insight emerges from experience obtained writing and applying software for generating random peptide structures. For the molecular dynamics experiments described in the next section, it appeared desirable to start each trajectory with a random conformation so as to eliminate a possible source of bias in sampling. Suitable random conformations cannot be generated merely by randomizing all dihedral rotations, because doing so would create knots and other impossible topologies that molecular dynamics might not relieve since doing so would require crossing bonds. Software was therefore written to take a peptide structure as input, randomize all dihedral rotations, and, proceeding outward from a randomly chosen starting point in the chain, iteratively perform dihedral rotations with backtracking until arriving at a conformation with no steric collisions.

Steric collisions were counted by identifying all non-bonded atom pairs where the distance between atoms was less than or equal to the sum of the van der Waals radii of each. A somewhat surprising observation was that it was necessary to reduce all of the van der Waals radii from the values typically assumed (by 0.22Ǻ, to 0.98Ǻ, 1.48Ǻ, 1.33Ǻ, 1.3Ǻ, and 1.63Ǻ for H, C, N, O, and S, respectively) for the collision removal algorithm to converge at all. Even with this relaxation of the collision criteria, the likelihood of a collision-free conformation resulting from the initial randomly chosen dihedral rotations is vanishingly small. Figure 2 shows a histogram of collision counts (number of pairs of atoms in collision) for the initial randomized conformations of a 20-mer peptide. Out of one million random conformations generated, none was completely free of steric collisions, and from the position and shape of the distribution of collision counts, it appears that the probability of obtaining a valid, collision-free conformation of a 20-mer peptide by assigning random rotations to all rotatable bonds is many orders of magnitude less than $10^{-6}$.

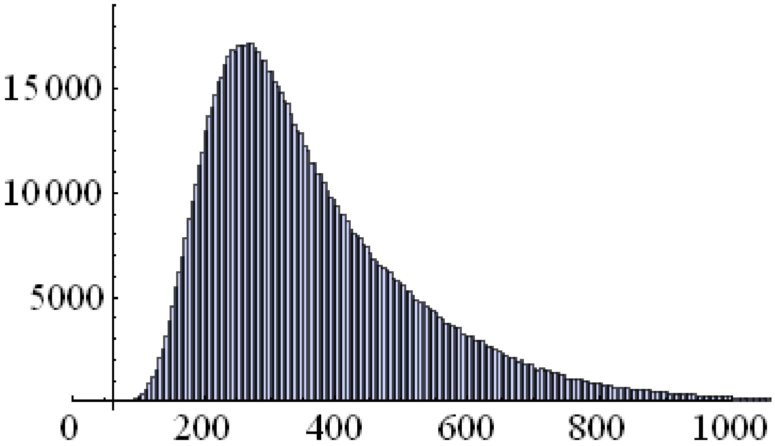

Figure 2. Histogram of atom-pair collision counts for 106 random conformations of a 20-mer peptide.

**Peptide conformational flexibility as observed in molecular dynamics experiments.**

To obtain a better understanding of the conformational diversity of CIM-10K library peptides, molecular dynamics simulations were performed on a 20-mer peptide selected via SPR experiments for high affinity to TNF-α (peptide TNF1, sequence FERDPLMMPWSFLQSRQGSC), and on an affinity-optimized variant of the same peptide with four substituted residues (peptide TNF1-opt, sequence FER**SY**L**K**MPW**K**FLQSRQGSC, substituted residues shown in bold). (The SPR screening was performed by Dr. Paul Belcher and Dr. Chris Diehnelt, and the affinity optimization was performed by Dr. Matthew P. Greving). These simulations provide further evidence in support of the conclusion that the random peptides here of interest are of intermediate conformational flexibility in the unbound state, each tending to spend much of their time in or near a handful of favored regions of conformation space, but also adopting many other shapes as they transition between attractors.

For each sequence, 100 molecular dynamics trajectories, each 10 ns in length, were generated using AMBER v.9 [131]. Each trajectory was begun from a conformation generated by assigning random values to all rotatable bonds, then iteratively rotating randomly chosen bonds to eliminate any steric collisions, then minimizing. Trajectories were run using a 2 fs time step, with AmberParm96 force field parameters, bonds to hydrogens constrained with SHAKE [132], and using the GB/SA implicit solvent model, with parameter settings SALTCON = 0.15, SURFTEN = 0.003, and EXTDIEL = 75 to simulate the salt, surfactant, and organic content of the SPR running buffer used for affinity measurements. Temperature for all runs was maintained at 300ºK via the Andersen thermostat [133] applied at 4 ps intervals. Conformations were sampled at 200 ps intervals after discarding the first 5 ns of each trajectory, yielding a total of 2600 samples for each sequence. A 2600 × 2600 pairwise distance matrix was computed reflecting average RMS distances following structural alignment of the backbone atoms of residues 4 through 11, as computed for each pair of conformations using Pymol's [134] “fit” function. Clustering was performed by iteratively identifying the largest subset of samples having RMS distances within a 1 Å threshold, and removing the cluster so identified from the distance matrix.

Figure 3 shows representative backbone conformations of the optimized region (residues 4 through 11) for each of the ten most highly populated conformation clusters for the TNF1 and TNF1-opt peptides, together with histograms showing the fraction of the 2,600 samples belonging to each cluster. (Conformations are shown with the N-terminal end at top, in descending order left to right by cluster size. The graphical representations of Figure 3 were produced using Pymol [134].)

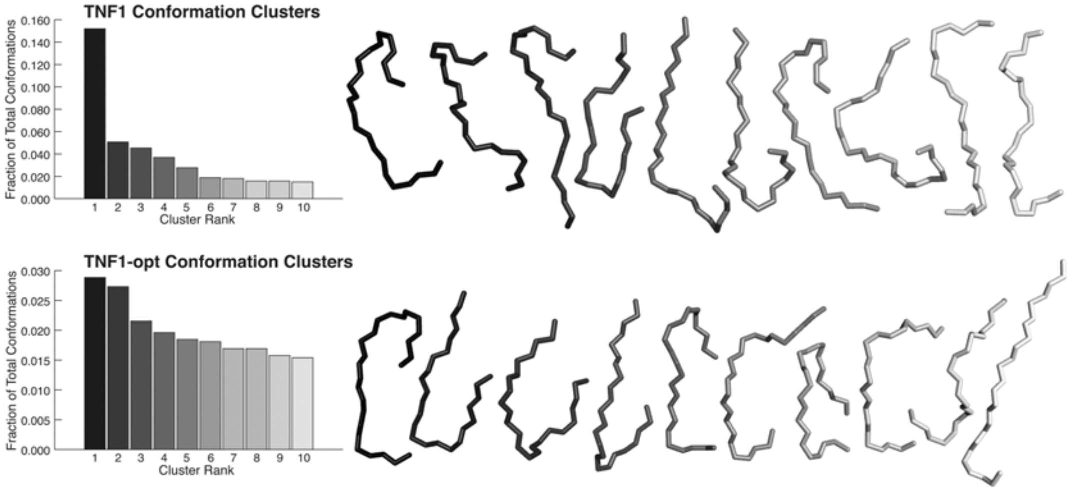

Figure 3. Molecular dynamics (MD) conformational analysis of the TNF1 (top) and TNF1-opt (bottom) peptides.

The 8 residue region of TNF1 from residue 4 through residue 11 is of particular interest because it is the region in which residue substitutions were found to improve the affinity of the peptide for TNF-α. In a total of 39.6 percent of the samples taken of TNF1, the residue 4-11 region adopts a conformation corresponding to one of the ten most populated clusters shown in Figure 3, with 15 percent of the samples conforming to the predominant cluster. Note that this tendency for the residue 4-11 region to favor certain conformations does not extend to the remainder of the molecule. Figure 4 shows the predominant cluster for TNF1; the aligned region can be seen in the center portion of the image, and forms a relatively tight cluster, while the distal regions show wide conformational diversity. The affinity-optimized variant, TNF1-opt, displays considerably greater conformational diversity than TNF1, with the predominant cluster accounting for only 2.8 percent of the samples, and the ten largest clusters comprising only 19 percent of the samples.  It is noteworthy that the mutations that best improved affinity appear also to have greatly increased conformational diversity; the implications of this observation as they relate to binding kinetics will be addressed in a later section.

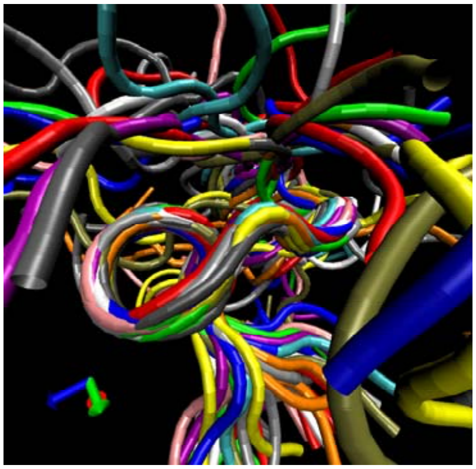

Figure 4. Predominant conformational cluster for TNF1, with backbone structure alignment on residues 4 through 11 and showing disordered ends.

Others have observed that “among secondary structure elements, β-turns are ubiquitous and major feature of bioactive peptides” [117], and that these may be important as recognition motifs [135]. The tendency to form β-turns, clearly seen in the conformations shown in Figure 3 and Figure 4, is typical of the behavior seen in shorter (typically 10 ns) molecular dynamics simulations performed on a number of other 20-mer sequences from the CIM-10K library. In repeated simulations, when a trajectory is begun from a relatively extended conformation, the peptide is typically seen to fold over on itself within about 5 ns or less, and transient hydrogen bonds form between the residues comprising the adjacent legs of the resulting hairpin. In these trajectories, the position of the β-turn is not fixed, but shifts along the middle region of the peptide as the legs of the hairpin adjust relative to each other, rapidly making and breaking hydrogen bonds as they do so. The first and last few residues of the chain typically extend outward from the hairpin structure and participate only transiently if at all in these interactions. Over the course of a 10 ns trajectory, no peptides have been observed to adopt a stationary conformation, and over longer trajectories, they are sometimes seen to return briefly to an extended conformation before again forming a β-turn.

Since protein folding occurs on a much longer time scale than 10 ns, these qualitative observations should not be over-interpreted, but the observed behavior is consistent with the hypothesis that polypeptides at physiologic temperatures in aqueous solvent should find it energetically favorable to adopt more compact conformations so as to minimize the exposure of hydrophobic moieties to the solvent. Since hydrophobic interactions, unlike hydrogen bonds, are not particularly fastidious in terms of alignment and identity of the interacting moieties, it is not surprising that the resulting structures tend to explore a wide range of positions. These observations will be addressed further in connection with the discussion of binding kinetics.

**Peptide polyspecificity in library screening experiments.**

If protein "shape space" is defined as the set of all possible distinctly recognizable combinations of shape, charge, and other relevant characteristics that binding regions on proteins can assume, most estimates of the size of that space would place it at ~$10^8$ or more distinct shapes [136]. (For economy of expression, "shape" will be used to refer to the totality of topological, physical, and chemical characteristics that determine the capability of a region of a biomolecule to participate in a specific binding interface with another biomolecule.) It has been estimated that immune recognition requires repertoire sizes of on the order of $10^7$ to $10^8$ for ~μM affinities, and up to $10^{10}$ for ~nM affinities [137]. Obviously, if each peptide were able to recognize only one "shape", there would be little likelihood of finding even one peptide capable of recognizing a given protein target from among a library of only 10,000 peptides.

In fact, in microarray experiments, most of the random 20-mer peptides in the CIM-10K library are significantly polyspecific, as is evident from a statistical analysis of the peptide microarray experiments performed (by others) in the Center for Innovations in Medicine over a several year period. In 1,322 separate microarray experiments involving a large number of distinct analytes ranging from peptides to proteins of various sizes and antibodies, with detection either by direct fluorescent labelling of the target or by fluorescent-labelled antibody specific for the target of interest, for each of the ~10,000 array peptides there were an average of 8 ± 3 targets for which fluorescent intensity was detected at five standard deviations or more above mean. In other words, each peptide showed considerably above background affinity for (on average) 0.6% ± 0.2% of the 1,322 analytes applied.  For an alternative perspective, for the same dataset of 1,322 array experiments, the average number of peptides that showed high binding (fluorescent intensity $\geq 5\sigma$ above mean) to the applied target in a single experiment was 57 ± 24 (0.6% ± 0.2%).

It would overreach the data to suggest that the foregoing array experiment statistics are directly comparable to the antibody repertoire-based estimates of the size of shape space described above.  A number of uncertainties may be cited. Target label intensities in array experiments give at best a rough estimate of affinities. Affinities corresponding to very high fluorescence intensities are typically in the low 10's of micromoles at best. Two array peptides that show affinity for a target may be binding different epitopes on the target. Target binding on microarrays may involve densely spotted peptides whose behavior arises from complex surfaces and/or avidity effects due to interactions of multiple copies of the peptide with each other or with the target.  Some of the array experiments involved polyclonal sera, with detection by anti-IgG secondary antibody, and in those experiments the applied analyte is obviously not homogeneous (although it is believed that the antibodies specific for the antigen vaccinated against substantially account for the observed response).

As a rough estimate, if it is assumed that there are $10^7$ distinct shapes recognizable by antibodies of moderate affinity, and the probability that an average random 20-mer peptide will recognize a given shape is on the order of 0.6%, then on average each peptide should be capable of recognizing $6 \times 10^4$ distinct shapes. The actual number is likely considerably smaller than that, due in part to the measurement uncertainties already described, and in part to the lower average affinity of peptides as compared to antibodies (since binding at high affinity should, on average, require more accurate shape complementarity, hence a larger shape space). Nevertheless, even taking all these reservations into account, if microarray experimental data bears any reasonable relationship to peptide affinities (and it seems to), it seems clear that each peptide is capable of binding at significant affinities to very many more than one target shape.

If it is assumed that array binding reflects, at least in significant part, interactions each involving a single peptide molecule and a single target molecule, then two possible general explanations for the apparent polyspecificity suggest themselves. The first is that the peptide might display multiple “epitopes”; even if the peptide assumes a single, relatively rigid structure, different parts of its exposed surface might be able to interact with different target shapes. For example, each three-residue subsequence of the peptide might be thought of as a separate entity capable of binding to an appropriate complementary shape. It is unlikely that rigid contacts of this kind would be sufficient by themselves to explain the observed polyspecificity, since, as will become apparent from the analysis of actual peptide-protein interfaces in Chapter 3, the energy of a single small contact would typically not be sufficient by itself to explain affinities of the magnitudes observed, and larger contacts would imply fewer distinct shapes. The other possible explanation is that each peptide is capable of adopting multiple conformations, so that it can accommodate a distribution of complementary shapes. These effects are not mutually exclusive, and it seems likely that the observed apparent ability of each peptide to recognize multiple target shapes is due to both conformational diversity and differences in which region(s) of the peptide surface are in contact with a particular target.

### Peptide structural diversity and the mechanism and kinetics of binding

The overall goal of the modeling described here is to inform the engineering design of peptide ligands. The mechanism and kinetics of peptide-protein binding are of interest because of their potential to inform the choice of structural and conformational characteristics that may best contribute to the desired binding behavior. It will be seen that affinities determined from single, presumed steady-state binding measurements are likely to be misleading if the actual binding mechanism deviates substantially from simple 1:1 Langmuir binding. The discussion to follow will begin with a brief review of the general theoretical principles underlying ligand-receptor binding kinetics. It will then draw from the literature on the kinetics of analogous interactions in search of the elements of a suitable model. After examining data obtained in surface plasmon resonance experiments (by others) on a CIM-10K library peptide, this section will conclude by suggesting the outlines of a hypothetical binding mechanism and discussing its implications as they relate to the selection of peptides for use as binding elements in synbodies.

### The kinetics of ligand-receptor binding

In simplest terms, binding of a ligand L to a receptor R may be described by the equilibrium expression:

$$R + L \xleftrightarrow[k_{OFF}]{k_{ON}} RL$$

The affinity of the ligand for the receptor may be expressed in terms of the dissociation constant, $K_D$ (or alternatively by $K_A$, the reciprocal of $K_D$) [53, 138]:

$$K_D = \frac{k_{OFF}}{k_{ON}} = \frac{[R][L]}{[RL]}$$

$K_D$ is expressed as a concentration, $k_{ON}$ is the rate constant for the forward (association) reaction in (here expressed in units of $M^{-1}sec^{-1}$), and $k_{OFF}$ is the rate constant for the reverse reaction (dissociation) in $sec^{-1}$. $K_D$ is nominally equal to $ED_{50}$, the ligand concentration at which the occupancy $\theta = 0.5$, meaning that 50% of the available receptor binding sites are occupied. (If $[L] = K_D$, it follows from the foregoing expression that $[R] = [RL]$ – that is, the concentration of unbound receptor equals the concentration of bound receptor.) This is, of course, true only at equilibrium, and others have noted that true $K_D$ may often be much less than experimentally measured $ED_{50}$ for slow $K_{ON}$ interactions where measurement is made before the system has reached equilibrium [49].

As already noted, the affinity is related to the Gibbs free energy of binding ΔG by the relation [49-51]:

$$\Delta G = RT\ln(K_D) = -RT\ln(K_A)$$

where R is the gas constant and T is the absolute temperature. It is important to keep in mind that the affinity, or equivalently ΔG, does not determine how rapidly equilibrium is reached. $K_D$ is a ratio of the reverse and forward rate constants $k_{OFF}$ and $k_{ON.}$ For example, $K_D = 1$ μM is consistent with $k_{ON} = 10^6\ M^{-1}sec^{-1}$ and $k_{OFF} = 1\ sec^{-1}$ (fast on, fast off), or $k_{ON} = 10^2\ M^{-1}sec^{-1}$ and $k_{OFF} = 10^{-4}\ sec^{-1}$ (slow on, slow off), or any other values that give the same ratio. Thus, from the standpoint of peptide design, affinity is not the only relevant variable; another is the speed at which the association reaches equilibrium, which is determined by the magnitudes of $k_{ON}$ and $k_{OFF}$. In particular, the time required to reach equilibrium may provide useful insight regarding the nature of the binding mechanism. It is also noteworthy that the speed of association and dissociation is a potentially engineerable quantity in itself, given an adequate understanding of how peptide and/or target properties affect the magnitudes of $k_{ON}$ and $k_{OFF}$, and, in the context of therapeutic applications, may have important implications for optimizing absorption, distribution, metabolism, and excretion (ADME) properties (e.g. [139]). The association/dissociation speed may also relate to specificity; since a relatively non-specific peptide can complex with a larger diversity of target loci, a relatively high association rate would be expected. Conversely, a highly specific interaction implies a very slow on rate because there are few ways in which it can occur.

The rate constant $k_{ON}$ specifies the rate at which new associations are formed. The rate of association may be approximated by a relation in the general form of the Arrhenius equation:

$$k_{ON} = Ae^{-E_a / RT}$$

where the “prefactor” A provides a measure of the collision frequency and the exponential term measures the probability that the collision results in a reaction. $E_a$ is the activation energy, corresponding in some models to the change in Gibbs free energy from the unbound state to the transition state ($\Delta G^{\ddagger}$). Thus, by way of illustration, a simple 1:1 interaction in gas phase can be modeled by:

$$k_a = \frac{k_B T}{h} e^{-\Delta G^{\ddagger} / RT}$$

where $k_a$ is (here) the reaction rate constant, $k_B$ is Boltzmann’s constant, T is the absolute temperature, h is Planck’s constant, and $\Delta G^{\ddagger}$ refers to the Gibbs free energy of the transition state. The exponential is simply the Boltzmann distribution based probability that a collision has energy greater than $\Delta G^{\ddagger}$.

Peptide binding to a protein in solution phase involves complexities that are not adequately captured by so simple a model. First, in solution phase the collision rate is greatly reduced due to solvent effects. This can be taken into account in part by approximating the collision frequency on the basis of an assumption that the association rate $k_{ON}$ is equivalent to the diffusion rate of the ligand through the solvent, given by $k_{diff} = 4\pi Da$ [51, 140] where D is the diffusion constant and *a* is the radius of the receptor or target. However, the implicit assumption of the diffusion-limited model that every collision will result in a stable association seems highly suspect as applied to structurally complex ligands like peptides, where many or most collisions may be wrongly oriented, occur between peptide and protein loci that are not capable of associating, and/or involve suboptimal conformations.

The "steric" problem – the requirement of optimal conformations and correct alignment of the colliding molecules – is sometimes dealt with via the concept of "reactive cross-section", which may be expressed as the product of a collision cross-section σ and a steric factor P in a modified Arrhenius-type expression such as:

$$k_a = P\sigma(4\pi Da)e^{-\Delta G^{\ddagger}/RT}$$

(See [50], equations 30.1.7, 30.1.5, and 26.2.5; [51], equation 18.26.) Even for interactions involving relatively simple molecules, the steric factor P is typically found to be much less than 1. In a textbook example (see [50], p. 739-40), for the gas phase reduction of ethylene to ethane ($H_2 + C_2H_4 \rightarrow C_2H_6$), the steric factor P is computed as $1.7 \times 10^{-6}$, and the author notes that "as a rough guide, the more complex the molecules, the smaller the value of P." It is hard to ascribe much predictive value to a model that requires a rate adjustment by six orders of magnitude on the basis of what amounts to a “fudge factor”. A further concern is that it is not necessarily clear how much, if at all, the conformational, orientation, and other constraints to which the steric factor P is addressed are already accounted for in the entropic component of the transition state energy $\Delta G^{\ddagger}$. Nevertheless, as will be seen in the discussion (below) of SPR data for the binding of a CIM-10K peptide to TNF-α, it may be possible to draw useful insight about the binding mechanism from a comparison of the estimated encounter rate with the observed association rate. Before turning to that analysis, however, it will be useful to review the literature and examine the binding models that have been proposed for similarly complex interactions.

### Binding models for interactions between complex molecules

Although ligand-receptor binding affinities are often expressed as though derived from simple 1:1 Langmuir equilibria, clearly such a model fails to capture fully even the vagaries of rigid ligand binding [141], much less all of the possible complexities of interactions involving peptides that exist in a distribution of conformations in the free solvated state, are potentially capable of conformational adaptations and transitions in the bound state, and may have a range of affinities with respect to multiple target sites. In the literature on macromolecular interactions can be found two general hypotheses either or both of which may contribute to an improved model [142]. The first is the “conformation selection” hypothesis [143, 144], in which binding would be assumed to require an encounter between an optimally oriented peptide molecule, already in or close to the conformation it will assume in the bound complex, and a binding site having complementary shape and charge. The second is the “induced fit” hypothesis [145-148], in which the peptide molecule interacts initially in some suboptimal manner, and thereafter changes in conformation, orientation, and/or position occur in the peptide, the protein binding site, or both, ending in a complementary bound conformation.

It will be argued here that both phenomena likely play a role in peptide binding.  The evidence for a diverse ensemble of peptide conformations has already been described; no great leap of logic is required to conclude that the probability that an encounter will result in an association event will vary depending on the peptide conformation and orientation at the time of the encounter. The discussion to follow will present the evidence, first from the literature and then from the limited experimental data available, that peptide-protein binding is likely also to involve formation of a temporary complex which then slowly transitions into a more stable, higher affinity final state.

The latter behavior can be described in simplified terms using a double-equilibrium model, first proposed in the context of protein-protein interactions [149-153], and later modified and extended to antibody-antigen binding [138, 154] and peptide-MHC binding [43].  Upon a productive collision of the ligand with the target an “encounter complex” is formed, energetically less favorable and therefore of lower affinity than the ultimate bound state, but adequate to maintain temporary contact. Over time, the complex then transitions to the final bound state:

$$A + B \xleftrightarrow[k_{-1}]{k_{+1}} AB^* \xleftrightarrow[k_{-2}]{k_{+2}} AB$$

and the overall equilibrium affinity constant is given by

$$K_A = K_{A1}(1 + K_{A2})$$

where $K_{A1} = k_1 / k_{-1}$ and $K_{A2} = k_2 / k_{-2}$ [138].

For antibody binding, Lipschultz, et al. [138] first demonstrated this multiphasic association in a anti-hen egg white lysozyme experimental system by showing in SPR experiments that dissociation rates decreased as association times were increased from 2 to 250 minutes. The minimum times required for the second reaction to progress to equimolar concentrations between the encounter complex and the final stable state – claimed to be a proxy for the $T_{1/2}$ of the transition from encounter to final state -- ranged up to 17 minutes. Kourentzi, et al. [154], analyzed the kinetics of a similar experimental system in greater detail; their results confirmed the two-step mechanism, and showed quite slow forward rates for the second equilibrium for some antibody-antigen combinations, with equilibration times on the order of hours, and $k_2 >> k_{-2}$. The mechanism implied by these findings was consistent with x-ray crystallographic evidence already in the literature indicating CDR loop rearrangements upon antigen binding in a specific antibody against a virus-derived peptide [147].

The encounter complex model is not, however, the last word in antibody binding kinetics. Though not as conformationally diverse as peptides, antibodies can exist in multiple unbound conformations, which is one factor contributing to their polyspecificity. In an anti-hapten antibody system, James, et al. [155] reported x-ray structures consistent with a conformation selection model, and kinetics measurements showed an equilibrium in solution between two distinct unbound forms, only one of which was capable of binding the antigen. Perhaps more relevant to peptide binding, Tsai et al. [144] have made a strong theoretical argument in favor of conformation selection as a model for interactions involving intrinsically disordered proteins. And Lange, et al., [143] have shown by NMR experiments on free unbound ubiquitin a measured ensemble that encompasses all of the binding site conformations found in 46 ubiquitin crystal structures, of which most involved bound complexes. Thus conformation selection cannot be ruled out, nor is it necessary to do so; as Grunberg, et al. [156] have suggested based on molecular dynamics and docking simulations, a model involving "a three –step mechanism of diffusion, free conformer selection, and refolding . . . is in better agreement with the current data on interaction forces, time scales, and kinetics."

Although there appears to be no comprehensive theoretical model of the binding kinetics of peptides specifically, useful direction can be obtained from studies examining peptide binding to specific targets. It will be seen that these are generally consistent with a multi-step model that takes into account diffusion, conformation selection, and a slow transition from encounter complex to final state.

Several groups have analyzed the kinetics of peptide binding to MHC class II [43, 49, 157, 158]. MHC binding is arguably somewhat unrepresentative of peptide-protein binding generally, in several respects. Because of the role that MHC presentation plays in immunity, it is essential that an MHC molecule be able to bind a large number of different peptide sequences, so there is an inherent bias in favor of polyspecificity. MHC-II binding is typically highly dependent on interactions with a few "anchor" residues [159], and tolerant of moderate peptide diversity outside those residues. Further, the MHC binding process must include mechanisms to ensure that presentation will not be dominated by a few of the best binding peptides [160]; perhaps for this reason, Kasson, et al. found that a variety of peptides bound MHC-II at approximately equal association rates [43].

From the MHC binding kinetics studies emerge several observations of potential relevance. First, the apparently very slow association rate for peptide-MHC-II binding was explained when it was discovered that unbound MHC-II exists in two or more forms, not all of which bind peptide [43, 161]. The inactive form, which predominates at equilibrium, converts to the active form at a very slow rate [43]. Hence, experimental measurements of $ED_{50}$ would tend to overestimate $K_D$, since part of the receptor population would be in the original state and unavailable for binding, and, worse, the proportion in a non-binding-capable state would be changing until the conversion reached equilibrium. For the peptide and MHC protein studied by Berezhkovskiy, et al., [49] the error ($ED_{50}/K_D$) was found to amount to as much as two orders of magnitude for a peptide with a very long dissociation half-life. This serves as a cautionary reminder about the importance of making affinity measurements at equilibrium. It also illustrates that when reactants exist in multiple forms, the association rate may depend not only on the proportion of binding-capable molecules in the ensemble, but also upon the rate at which non-binding-capable molecules convert. Thus for peptides in solution in a distribution of conformations, it would be expected on the basis of Le Chatelier's principle that the removal of the binding-capable conformations by binding to the protein target would cause a net transformation of other conformations to the binding-capable form, considerably extending the time required to reach equilibrium (cf. [144]).

Second, as noted, the range of association rates found by Kasson, et al., for 20 distinct peptides was quite narrow ($4.4 \times 10^4$ to $4.0 \times 10^5$ $M^{-1}sec^{-1}$, mean $1.45 \times 10^5$ $M^{-1}sec^{-1}$) "despite having dissociation rate constants that span a range of greater than 10,000-fold". Notwithstanding the reservations about the unique aspects of MHC binding, this data provides at least a point of reference for evaluating estimated on rates. Also, since the sequences in question range from 8 to 16 residues in length, by the usual method of estimation (a fixed penalty per rotatable bond) they would be expected to span a wide range of conformational entropy. Yet the narrow range of association rates implies that the transition state energies ($\Delta G^{\ddagger}$) must also lie in a relatively narrow range. This in turn suggests that the entropic component of $\Delta G^{\ddagger}$ must be relatively unimportant, meaning that the peptide must remain relatively flexible and unconstrained in the transition state, and transition over time to a more tightly bound final state corresponding to the observed dissociation half-times. Inferential support for this (admittedly speculative) hypothesis may also be found in the conclusion reached by Kasson, et al. that the transition state depends on relatively nonspecific hydrophobic interactions and "does not depend greatly on interactions between the protein and side chains of the peptide." [43]. This behavior is consistent with the hypothetical binding model proposed in the concluding section of this chapter.

The study by Goldberg, et al. [162] of the association kinetics of a modified RNase protein with a 15-residue peptide lends further support to the foregoing inferences. Here, the side chains of two residues of the peptide were found to contribute most of the affinity and accounted for a large fraction of the overall interface. Substitution at these two positions caused up to a 6 order of magnitude loss of affinity, while leaving $k_{ON}$ essentially unchanged at about $10^7$ $M^{-1}sec^{-1}$ ([162], figure 1), again suggesting that the association mechanism is not particularly specific or dependent on native-like interactions between peptide side chains and the protein. The unusually high on rate arguably supports this conclusion, since it implies that an unusually high percentage of collisions result in an association event.

Takeda, et al. [163] measured the kinetics of binding of a 7-residue peptide to a heat shock protein (Hsc70) in the presence of either MgADP or

MgATP. The ratio of ATP to ADP declines under stress conditions, and is thought to regulate the chaperone activity of the heat shock protein, which has a relatively high affinity (here $K_D = 4.3$ μM) for peptide in the presence of MgADP and a low affinity in the presence of MgATP ($K_D = 40$ to 50 μM). The high affinity binding in the presence of MgADP is characterized by two-step kinetics [164], with an initial rapid association ($k_{+1} = 1.21 \times 10^3$ $M^{-1}sec^{-1}$) to a lower affinity bound state ($K_{D1} = 14.2$ μM), followed by a much slower transition ($k_{+2}$ = 0.013 $sec^{-1}$) to a high affinity final state ($K_2 = 0.29$):

$$Hsc70 + P \xleftrightarrow[k_{-1}]{k_{+1}} Hsc70P \xleftrightarrow[k_{-2}]{k_{+2}} (Hsc70P)^*$$

(In the presence of MgATP, the heat shock protein undergoes a conformational change, preventing the transition to a high affinity state, and resulting in single-step kinetics corresponding closely to the initial rapid but low affinity association.) As with MHC binding, chaperones are required to bind a diverse selection of disordered polypeptide substrates, so the low specificity may make heat shock proteins a poor basis for a general binding model. Nevertheless, the Hsc70 data offer yet another example of a peptide-protein binding mechanism characterized by rapid formation of an initial complex, followed by slow transition to a high affinity final state. Moreover, the initial complex again inherently involves a low specificity, non-fastidious interface.

Multi-step kinetics were also observed in a study of the interaction of a 40-residue amyloid-β (Aβ) peptide to an affibody protein ($Z_{A\beta3}$) (a 15.6 kD cysteine-linked homodimer having four helical domains, with certain residue positions variable, here selected for high affinity to Aβ) [165]. It was found that the bound $Z_{A\beta 3}$:Aβ complex has conformational characteristics not found in the free affibody or peptide: the C-terminal part of the Aβ peptide (residues 17-36) adopts a β-hairpin conformation and 5-residue regions of each of the $Z_{A\beta 3}$ affibody subunits form β-strands. Analysis of calorimetry data indicated that these conformation changes occur during a slow (relaxation time = 13 sec) transition phase following rapid formation of a quite high affinity (1 μM $K_D$ or better) encounter complex.

It is apparent that in essentially all of the studies found in the literature involving the kinetics of interactions reasonably analogous to peptide-protein interactions, it was found necessary to incorporate in the binding model an induced fit / encounter complex mechanism, conformation selection, or both in order to obtain realistic measurements of the kinetic properties upon which affinity estimates would be based. It will be shown in a later section that the available data relating to CIM-10K library peptides, though limited, supports a similar conclusion.

### Multiplicity and non-homogeneity of contact regions

When a peptide and a protein collide, the encounter results in a contact interface, perhaps transient, in which part of the peptide comes into contact with part of the protein surface. (More accurately, there is a set of peptide atoms each of which is closer than some threshold interaction distance to at least one atom of the protein, and vice versa.) Even for a relatively small protein, the number of possible distinct surface patches of appropriate size for interaction with a peptide is enormous. Many of these surface patches will be capable, at least in principle, of being bound at least weakly by a suitable ligand. A study of the diversity of a B cell response to tetanus toxoid estimated that the resulting polyclonal antibody repertoire represented on the order of 100 distinct clonal selection events [166], suggesting that the tetanus toxoid antigen, at least, exposes many loci potentially recognizable by antibodies. Others have estimated that a typical repertoire of antibody paratopes can recognize on the order of 10 distinct epitopes on a typical 50 kD protein [167].

When the conformational diversity of the peptide is taken into account, as well as the range of peptide surface regions that can be presented at various rotations / translations, it is obvious that if an ensemble of contact interfaces could be constructed by sampling a large number of random encounters, that ensemble would encompass a huge range of distinct contact interfaces. In many of the interfaces in such an ensemble, only a part of the peptide chain will be in contact with the protein. As will be shown in Chapter 3, even in stable bound interfaces, the enthalpy of binding appears to be distributed quite non-uniformly over the peptide contacts, with a disproportionate contribution attributable to a few “hot spot” regions. It is therefore germane to inquire whether it may be advantageous to model the peptide as an aggregate of individually interacting fragments.

Fragment-based methods are attracting increasing interest, particularly in the field of small molecule drug discovery, where they are seen as a way to search greater expanses of chemical space while focusing on combinations of moieties most likely to contribute the desired binding properties. (E.g. [168-173] and the recent review by Hajduk and Greer [174]). Lead candidates may be constructed by joining relatively low affinity fragments (essentially the strategy underlying the synbody platform), and/or by improving the affinity of a fragment by functional group addition or alteration. As described in a previous section, many protein structure prediction algorithms base their estimates in part on computed or statistically sampled structures of short peptide fragments (e.g. [101]), but no reports have been found of fragment based design or discovery of intermediate length peptide ligands. To extend the concept to peptide ligand design, it would be logical to employ computational docking or QSAR methods to identify small peptidic fragments of high ligand efficiency ($\Delta G_{bind}$ per unit size or molecular weight), construct a combinatorial library from the fragments so identified, and select the best performers. The main obstacle to such a strategy is the final selection step; for reasons explained in detail in Chapter 4, molecular docking of full length 20-mer peptides is computationally intractable and chemical synthesis and testing of the requisite number of peptides is prohibitively expensive. It is, however, possible to approach the problem probabilistically, and estimate the behavior of the combined full length peptide by reference to the spatial relationships among the binding loci preferred by the fragments and the fragment binding energies at those loci. In Chapter 4, a strategy along these lines is shown to have merit for predicting binding sites of full length peptides. The probabilistic approach has another advantage: it allows for the possibility that each fragment may have affinity for more than one protein surface locus. Several small, higher affinity fragments of a single highly flexible peptide might be capable of binding more than one combination of loci, so that the peptide as a whole might be capable of binding in multiple positions, as illustrated in Figure 5. More than one possible bound position could also occur if two or more larger fragments, each able by itself to supply the entire binding energy required for the peptide to bind, are each complementary to a different locus on the protein, or if a single larger fragment is able to adopt alternative conformations each complementary to a different locus.

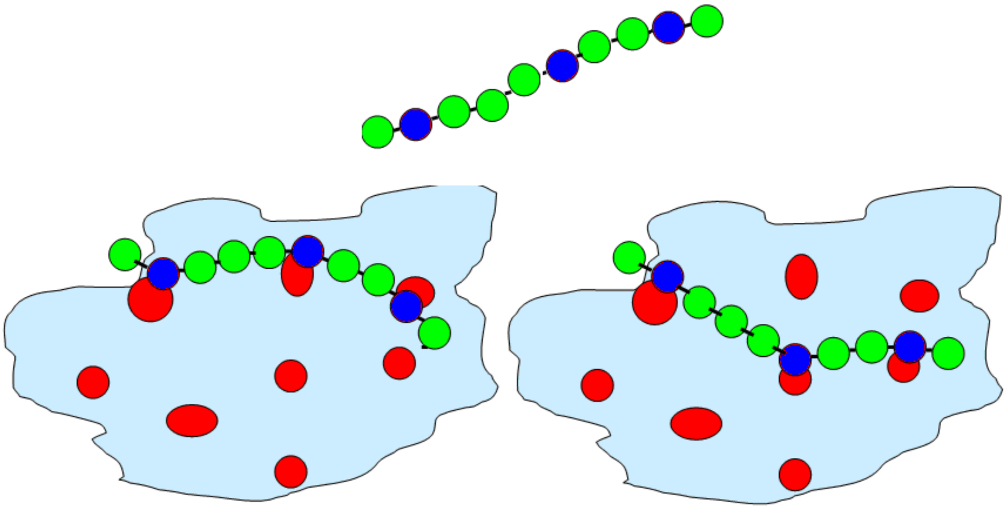

Figure 5. Alternate binding positions from fragment binding at multiple loci.

The theta state-related polymer properties of polypeptides provide an interesting perspective on the question of what fragment sizes are most informative. As discussed in a previous section, the degree of polymer chain expansion or collapse depends on the temperature and solvent conditions, with temperatures above the theta temperature favoring expanded states. The dynamics of this behavior varies, however, depending on the length scale. For a segment of the chain to collapse against another segment, the magnitude of the forces mediating the collapse must be on the order of $k_BT$ or above, and a very short segment cannot interact over a sufficient area for the attractive forces to sum to the requisite energy [62]. Therefore, according to "blob theory" [62, 63, 175], polymer chain expansion and collapse occurs at the scale of "blobs", which are fragments of the maximum size such that the balance of forces within the fragment remains below the "order of $k_BT$" threshold. Pappu et al. have estimated that for a range of protein sequences, the "blob" scale is about 7 residues [62, 175]. This estimate meshes rather well with the data on the TNF1 and TNF1-opt peptides discussed above in connection with Figure 3 and Figure 4. Recall that all of the four substitutions that were found (by others) to best optimize the affinity of the TNF1 peptide for TNF-α were within the 8-residue segment from residues 4 to 11. Moreover, as is clearly seen in Figure 4 for the predominant cluster (and was also true for other clusters), in molecular dynamics simulations it was possible to find in the sampled ensemble tight clusters of conformations of the same 8-residue region, in which the conformations of the remainder of the peptide were dissimilar. These data arguably support a hypothesis that the 8-residue fragment is acting as a relatively independent subunit whose binding properties can be analyzed as a separate fragment.

This does not necessarily mean that fragment interactions at a smaller length scale are irrelevant. Arguably, the size of (relatively) independently binding fragments relates to the issue of specificity. By definition, an optimally specific ligand would bind strongly to a single locus on a single target and bind very weakly or not at all to any other substrate. Other factors being equal, high structural diversity implies low specificity -- the greater the diversity of shapes that a ligand can adopt, the greater is the diversity of potentially complementary target loci. At one extreme, a peptide in which each residue behaves independently of each other residue will be capable of high conformational diversity, and likely exhibit a high degree of non-specific binding. At the other extreme would be a rigid ligand; the highly specific protein-protein binding upon which many regulatory processes depend is enabled by binding sites that are held relatively rigid by the folded structure underlying them. A peptide in which the independently acting units are ~7-residue blobs, where the blobs are each capable of a limited repertoire of preferred conformations, may represent a practicable compromise. From the peptide microarray data already presented, it appears likely that most CIM-10K library peptides inhabit the higher structural diversity end of the spectrum.

### Association kinetics of CIM-10K library peptides

Equilibration behavior observed in peptide microarray binding tends to confirm the hypothesis of slow transition to a final equilibrium. In experiments by Rebecca Halperin on robotically spotted peptide microarrays each having approximately 10,000 distinct features corresponding to the CIM-10K library, fluorescently labelled monoclonal anti-P53 was applied generally in accordance with the protocol described in [176], with fluorescence intensities measured (on separate arrays) immediately after application, and after incubation times of 18

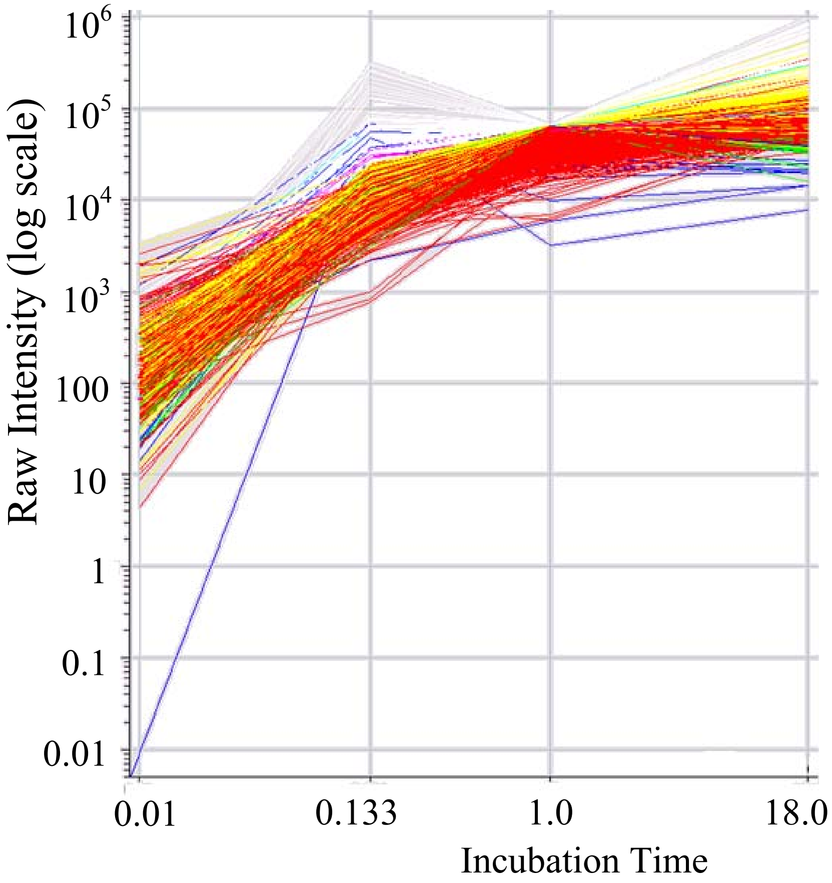

Figure 6. Array intensity vs. incubation time.

minutes, one hour, and 18 hours. Figure 6 shows raw fluorescence intensities measured for peptides registering above an arbitrary intensity threshold. Increasing the incubation time from 18 minutes to one hour increases the intensities corresponding to these higher-binding peptides by approximately an order of magnitude; the intensities rise further as the incubation time is lengthened to 18 hours.

These data are obviously far from conclusive regarding the kinetics of antibody binding to the array peptides. Others have observed that although peptide microarray data is useful for distinguishing “good binders” from “non-binders”, the error range for array-based affinity measurement may be quite high [177]. And, reservations about accuracy aside, Figure 6 merely shows slow equilibration, which could be caused either by a conformation selection mechanism or by a slow transition to final state after initial encounter, or both, or by other confounding factors such as aggregation.

Available SPR data on CIM-10K library peptides is not ideal for examining equilibration rates because the association times used in SPR assays were short. The sensorgram shown in Figure 7 is typical of SPR data for binding of peptides selected from the CIM-10K library, in solution phase, as “good” binders to surface-affixed target proteins. Here, in an experiment conducted by Dr. Paul Belcher and Dr. Chris Diehnelt, peptide TNF-1 (sequence FERDPLMMPWSFLQSRQGSC) was flowed at concentrations ranging from 800 nM to 27 μM over TNF-α affixed to the gold surface of the SPR chip using a Biacore A-100 SPR instrument. (See [33] for experimental details.)

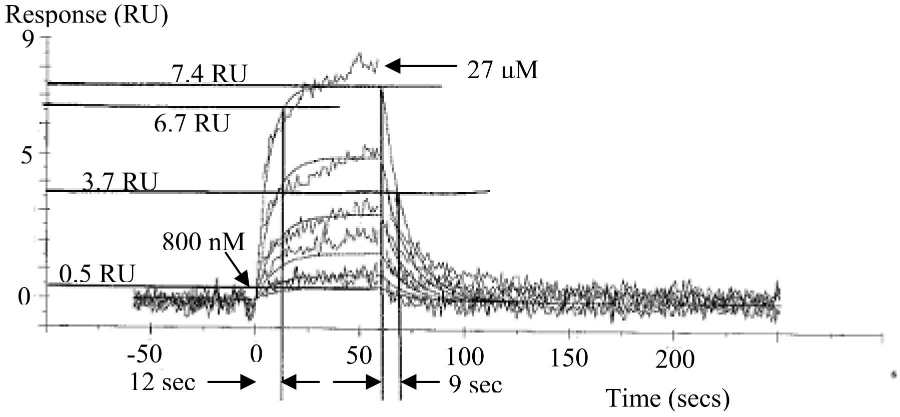

Figure 7. SPR sensorgram applying peptide TNF1 to TNF-α

Assuming simple 1:1 Langmuir kinetics, $k_{ON}$ and $k_{OFF}$ can be estimated from the sensorgram, giving results that are in close agreement with the affinity ($K_D$ of approximately $2 \times 10^{-5}$) estimated by the Biacore software. $k_{OFF}$ is easily computed from the observed 9 second dissociation half-time (the time required after the beginning of the dissociation phase for the response to drop by 50%, from 7.4 RU to 3.7 RU, referring to the 27μM curve). Since the decay is exponential,

$$k_{OFF} = \ln(2)/\mathrm{T}_{1/2} = 0.69302/\mathrm{T}_{1/2} = 0.077\sec^{-1}$$

Computation of $k_{ON}$ is less straightforward, since during the association phase both association and dissociation are occurring, and dissociation increases as occupancy increases. To compute $k_{ON}$, the time constant $k_{ob}$ for the (assumed) exponential rise to the steady state plateau level $Y_{\max}$ may be computed by fitting the data to:

$$Y = Y_{\max}(1 - e^{-k_{ob}t})$$

Using the foregoing relation, an approximation of $k_{ob}$ can be obtained from the time required for the response to rise from zero to (say) 90% of $Y_{\max}$, here approximately 12 seconds as shown in Figure 7:

$$0.9 = 1 - e^{-k_{ob}t}$$

$$k_{ob} = \frac{-\ln(0.1)}{12\sec} = 0.192\sec^{-1}$$

Given the applied peptide concentration [P] and the already estimated $k_{OFF}$, $k_{ON}$ and $K_D$ can be computed:

$$k_{ON} = \frac{k_{ob} - k_{OFF}}{[P]} = \frac{0.192\sec^{-1} - 0.077\sec^{-1}}{27x10^{-6}M} = 4.26x10^{3}M^{-1}\sec^{-1}$$

$$K_D = \frac{k_{OFF}}{k_{ON}} = 1.8x10^{-5}M$$

From the measured $K_D$ of 18 µM and the applied peptide concentration [P] of 27 µM, the occupancy can be computed:

$$\theta = \frac{K_A[P]}{1 + K_A[P]} = 0.6$$

It should be obvious from the SPR sensorgram (Figure 7) and from the known characteristics of the interacting species that a simple kinetics model is unlikely to describe accurately the mechanism of at work here. Two observations (most easily seen in the 27 µM trace) are worthy of note: (1) the SPR response appears to be continuing to rise at the end of the 60 second association phase, although the smoothed association curve appears to flatten (solid lines, fitted by software provided by Biacore, using an unknown algorithm); and (2) in the 190 second dissociation phase shown, during which the peptide concentration in the applied buffer was zero, the response does not return to zero.

Assuming they are real and not due to measurement error, nonspecific binding, aggregation, or other artifacts, these observations can be explained by either of two general mechanisms.

First, the peptide may be capable of binding to more than one site, at different affinities. Suppose, for a greatly oversimplified example, that the peptide has a high affinity of 1.6 μM for its preferred binding site, with $k_{+1} = 25$ $M^{-1}sec^{-1}$ and $k_{-1} = 0.00004$ $sec^{-1}$. Suppose further that the peptide is also capable of relatively weak and non-specific binding to other sites, with an average affinity of 100 μM, corresponding to $k_{+2} = 800$ $M^{-1}sec^{-1}$ and $k_{-2} = 0.08$ $sec^{-1}$. If it is assumed that these binding mechanisms are independent, the system is equivalent to:

$$P + T_1 \xleftrightarrow[k_{-1}]{k_{+1}} C_1$$
$$P + T_2 \xleftrightarrow[k_{-2}]{k_{+2}} C_2$$

where $C_1$ represents the the peptide in complex with the high affinity site and $C_2$ the complex with the low affinity site(s). See [178]. Representing the concentration of the target protein by T, the concentration of the preferred binding site by $T_1$, the concentration of the weaker binding sites by $T_2$, and the concentrations of peptide and the bound complexes by P, $C_1$, and $C_2$, respectively, the kinetics are described by:

$$
\begin{aligned}
T_1' &= -k_{+1}PT_1 + k_{-1}C_1 \\
T_2' &= -k_{+2}PT_2 + k_{-2}C_2 \\
C_1' &= +k_{+1}PT_1 - k_{-1}C_1 \\
C_2' &= +k_{+2}PT_2 - k_{-2}C_2
\end{aligned}
$$

where $T_1$', $T_2$', $C_1$', and $C_2$' are the time derivatives of the respective concentrations. To simulate an SPR experiment, the peptide concentration P is taken as constant during the association phase, and as zero during the dissociation phase. For starting concentrations $T_1$ = 100 µM, $T_2$ = 300 µM, P = 27 µM, and rates $k_{+1}$ = 25 $M^{-1}sec^{-1}$, $k_{-1}$ = 0.00004 $sec^{-1}$, $k_{+2}$ = 800 $M^{-1}sec^{-1}$, $k_{-2}$ = 0.08 $sec^{-1}$, the foregoing system was iterated by finite differences, giving the behavior shown in Figure 8.

On the assumptions stated, the affinity of the peptide for the preferred binding site is $K_D$ = 1.6 µM, and for the low affinity binding, $K_D$ = 100 µM; there are assumed to be three low affinity sites and one high affinity site per molecule. The resulting kinetics are essentially indistinguishable from the actual experimental data of Figure 7. The low affinity, fast on / fast off binding to suboptimal sites swamps the measured response to produce an affinity estimate, ~18 µM, calculated as from Figure 7 on the basis of 1:1 kinetics, that is more than an order of magnitude worse than the “true” affinity.

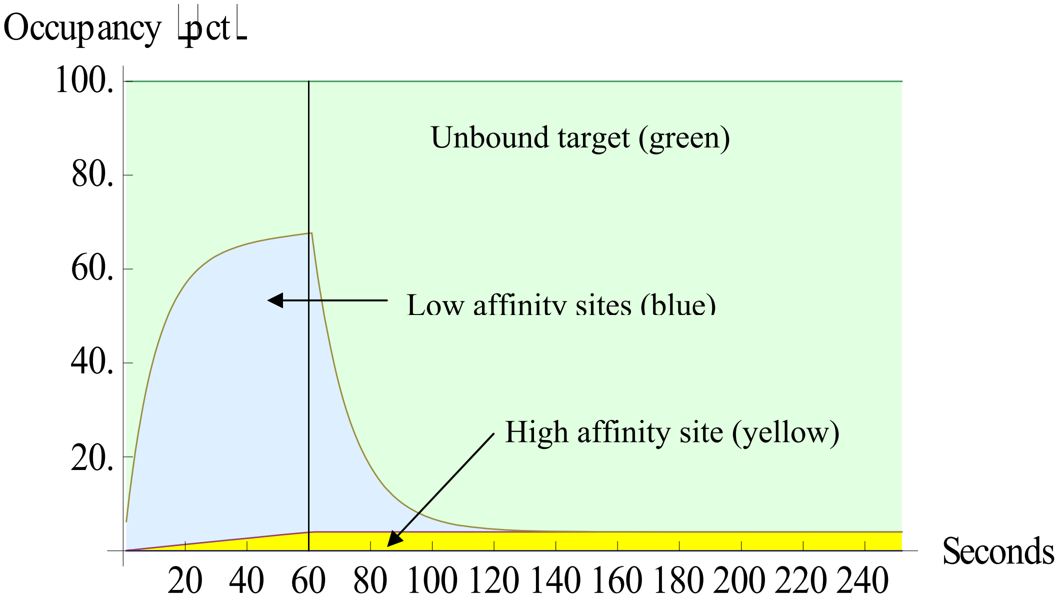

Figure 8. SPR simulation including low affinity non-specific binding.

The second mechanism that could explain the observed sensorgram data is a sequential equilibrium of the kind already discussed:

$$P+T \underset{k_{-1}}{\overset{k_{+1}}{\longleftrightarrow}} C_1 \underset{k_{-2}}{\overset{k_{+2}}{\longleftrightarrow}} C_2$$

where the peptide P and target protein T initially form an encounter complex $C_1$, and the encounter complex undergoes a change in conformation, position, or both to arrive at a final state $C_2$. The kinetics are described by:

$$T' = -k_{+1}PT \ + k_{-1}C_1$$
$$C_1' = +k_{+1}PT_1 - k_{-1}C_1 - k_{+2}C_1 + k_{-2}C_2$$
$$C_2' = +k_{+2}C_1 - k_{-2}C_2$$

where P is the peptide concentration, again held constant during the association phase and set to zero during disassociation, T is the target protein concentration,

$C_1$ and $C_2$ are the concentrations of the encounter complex and final bound complex, respectively, and T', $C_1$', and $C_2$' are the time derivatives of T, $C_1$, and $C_2$. Figure 9 shows the simulated response for assumed starting concentrations of T = 100 μM, P = 27 μM, and $C_1 = C_2 = 0$, with rates $k_{+1} = 4100\ M^{-1}sec^{-1}$, $k_{-1} = 0.077\ sec^{-1}$, $k_{+2} = 0.00173\ sec^{-1}$, and $k_{-2} = 0.00231\ sec^{-1}$, corresponding to relatively rapid formation of the encounter complex, followed by a slow transition to the final bound state. Again, the response data is identical in practical terms to the measured response shown in Figure 7, but if the process is allowed to reach equilibrium (requiring ~30 minutes), the affinity indicated by the total occupancy at equilibrium (θ = 0.72) is 10.7 μM.

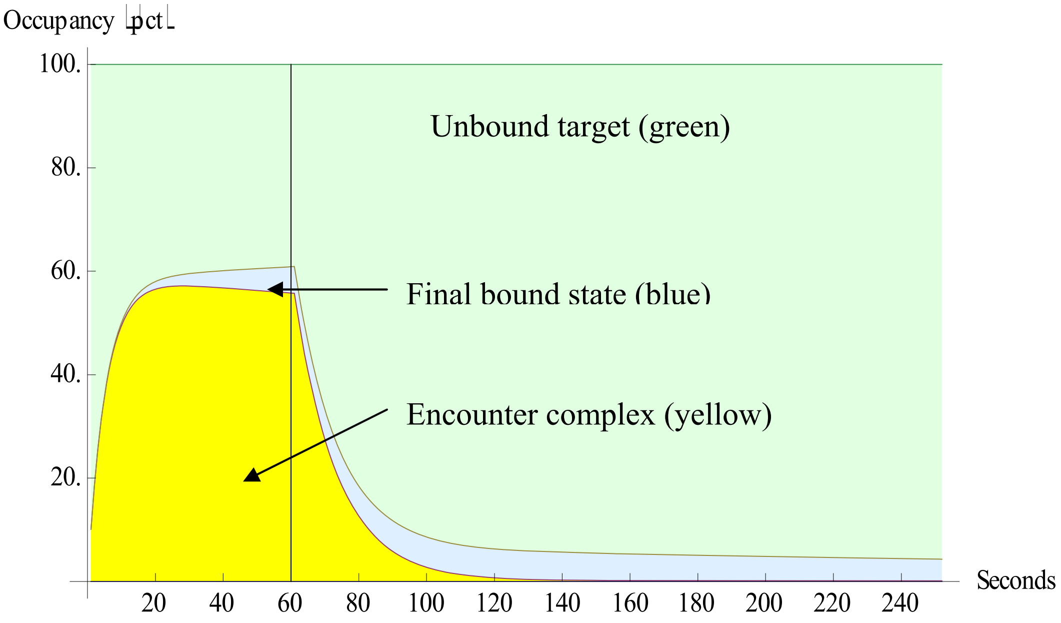

Figure 9. SPR simulation for encounter / transition model.

Conformation selection could be a factor in either of the foregoing mechanisms. For an SPR-type experiment in which the concentration of peptide is held constant during the association phase, conformation selection can be simulated by reducing the assumed peptide concentration to that fraction of the total concentration representing the binding-capable conformation. From the molecular dynamics sampling already described, it appears likely that that fraction may be quite small; even the largest conformation cluster for the optimized TNF1-opt peptide in Figure 3 comprised only 2.8% of the ensemble, and there is no reason to suppose that the largest cluster necessarily corresponds to the binding-capable conformation. If, for example, it is assumed that only 1 percent of the ensemble comprises conformations that are capable of binding, then to obtain an SPR response equivalent to that for a 1:1 binding mechanism with $K_D$ = 18 μM and [P] = 27 μM, as approximated by Figure 7, it would be necessary to assume a 100-fold increase in the association rate for that conformation, implying $k_{ON} = 4.26 \times 10^5$ and a "true" affinity of 0.18 μM. This line of reasoning is of interest in part for its potential to allow better estimation of the entropic component of ΔG. As will be seen in Chapter 3, it is feasible, albeit computationally nontrivial, to estimate ΔH from force field or other factors, given the exact geometry of a peptide-protein interface. ΔS cannot be determined from the structure of the bound complex. If, however, experiments could be devised that would report the association constant of the binding-capable conformation alone, then the value of ΔG computed from that association constant would be entirely due to ΔH except for the 6 degrees of rotational and translational freedom of the peptide relative to the protein (assuming the protein is rigid), and that entropic penalty would be the same in all cases. A strategy to place bounds on the binding-capable fraction of the ensemble might be to evaluate the extent to which the bound conformations of peptides in known interfaces correspond to conformations in a free liquid phase ensemble computed by molecular dynamics methods. A database of peptide-protein interface structures has already been assembled in connection with the work described in Chaper 3, and methods for sampling and clustering the free peptide ensemble were explored as previously described in this chapter, so performing the analysis should be relatively straightforward, if demanding in terms of computational resources.

It would be also be useful to know the relative importance of conformation selection and post-encounter transition for purposes of better optimizing synbody peptides. If conformation selection is the dominant mechanism in peptide binding, the optimal strategy would obviously be to design peptides that strongly favor the binding-capable conformation. If binding depends mainly on a post-encounter transition, however, a better candidate might be a peptide that forms non-specific encounter complexes easily and is flexible enough to “hunt” for a high affinity final state. The two mechanisms may also have different implications regarding multivalent binding: if peptide binding were entirely a matter of conformation selection, then the association rate for a bivalent interaction should be much less than that of individual peptides, since the probability of both peptides being in binding-capable conformations would be the product of the probabilities for each peptide separately, assuming independence.

An upper bound on the encounter rate can be estimated based on diffusion. The diffusion constant D depends on the size and structure of the diffusing molecule, and on solvent properties and temperature. For a molecule of the size of a peptide in aqueous solution, the possible range of values is small; the reported diffusion coefficient measured at 298°K and 0.5 mM concentration for a 16-mer peptide of $D_{peptide} = 2.37 \times 10^{-6}$ cm$^2$/sec ([179], figure 3) furnishes a reasonable basis for an estimate. The TNF-α homotrimer has the shape of a truncated cone about 6 nm in height and 5.5 nm across the base [180, 181], so the radius may be estimated at approximately 3 nm. Its diffusion coefficient may be approximated as $D_{protein} = 1.1 \times 10^{-6}$ cm$^2$/sec, as measured for lysozyme, a protein of similar size and shape to TNF-α [182]. Then

$$\begin{aligned}&k_{diff} = 4\pi a(D_{peptide} + D_{protein}) = 4\pi \cdot 3\times10^{-9}m \cdot 3.47\times10^{-6}cm^2/\sec\cdot 10^{-4}m^2/cm^2\\&= 1.31\times10^{-17}m^3/\sec\end{aligned}$$

[140]. The diffusion-controlled collision rate per molecule of target is the product of $k_{diff}$ and the concentration:

$$\begin{aligned}&I_{diff} = -k_{diff}[P] = 1.31\times10^{-17}m^3/\sec\cdot 27\times10^{-6}mol/L\cdot10^3L/m^3\cdot 6.02\times10^{23}/mol\\&= 2.13\times10^5\sec^{-1}\end{aligned}$$

The number of encounter complex-forming events per second required to produce the observed on rate is given by

$$I_a = k_{ON}[P] = 4.26x10^3 M^{-1}\sec^{-1}\cdot 27\times 10^{-6} M \\ = 0.11 \quad \sec^{-1}$$

Therefore the diffusion controlled rate is a factor of 1.93 × $10^6$ faster than the observed association rate. It must be emphasized that this merely establishes a bound; it is not valid to conclude that the actual collision rate is 213,000 per second, because the derivation of the expression for the diffusion controlled rate is based on an assumption that all ligand molecules bind immediately upon collision, producing a ligand concentration of zero at the target surface, with diffusion toward the target driven by the resulting concentration gradient (see [51, 140]). Since it is apparent that only a very small fraction of peptide-protein encounters result in formation of a bound complex, presumably the concentration gradient is greatly reduced.

An alternate method here proposed for estimating whether the interaction dynamics provide adequate opportunity for conformation selection is to compute what might be termed the average dwell time; that is, how much time, on average, each target molecule spends in close proximity to a peptide molecule before a complex is formed. This can be estimated using a model in which a unit volume is populated by a quantity of target molecules equivalent to the target concentration of interest, placed at random positions within the unit cube (with steric overlap, i.e. centers closer than the sum of molecular radii, disallowed). The volume is further populated by a quantity of ligand molecules corresponding to the ligand concentration, again placed randomly and with steric overlap disallowed. It is then a simple matter to determine the percentage of target molecules that are closer than an arbitrary interaction threshold distance to a ligand molecule. This model may also be more realistic in terms of the hypothesized dynamics than a diffusion-driven collision model, in the context of interactions taking place in a solvent between large, slow moving, relatively flexible molecules, where solvent caging effects (see [140] at 226; [183, 184]) may tend to keep interacting pairs from separating rapidly after an unproductive encounter.

Again assuming a peptide concentration of 27 μM, a cube 1000 nm on a side would contain $N_{pep} = 27 \times 10^{-6} \frac{M}{L} \cdot 10^{-15} \frac{L}{\mu m^3} \cdot 6.02 \times 10^{23} M^{-1} = 16{,}254$ peptide molecules. If peptides are modeled as spheres of radius 1.5 nm, and the target protein molecules are modeled as spheres of radius 3 nm, the volume from which the centers of target molecules must be excluded is $V_{pep} = \frac{4}{3}\pi(4.5nm)^3 \cdot 16254 = 6.2 \times 10^6$ nm$^3$. A reasonable distance defining the range at which a peptide and protein molecule might be deemed to be interacting is given by the the Bjerrum length, the distance at which the energy of electrostatic attraction of two opposite single charges is equal to the thermal energy RT; this is about 7Å in water at physiological temperatures [51]. Then the volume surrounding a peptide within which the presence of the center of a target molecule could be said to represent an interaction is

$V_{\text{int}} = \frac{4}{3}\pi(5.2nm)^3 \cdot 16254 - V_{pep} = 3.4 \times 10^6\,\text{nm}^3$. That volume represents $3.4 \times 10^6 / (1 \times 10^9 - 6.2 \times 10^6) = 0.34\%$ of the volume not occupied by peptide, suggesting that, on average, 55 of the target molecules (0.34% of 16,254) are within the Bjerrum distance of a peptide at any given time.

Based on the observed association time constant $k_{ob} = 0.192$ for binding of peptide TNF1 at 27 μM concentration to TNF-α, the time required to reach 10 percent target occupancy is 0.95 second. Therefore, in the first second, 1,711 of the 16,254 target molecules in the 1000 nm cube would be expected to have formed complexes with peptide. Then the average dwell time required for a complex to result is 55/1,711 = 0.032 seconds, abundant time to allow conformation selection or post-encounter transition or both. (By way of comparison, recall that the distribution of peptide conformations shown in Figure 3 was obtained by sampling from a 0.000001 sec molecular dynamics trajectory, and appeared to span reasonably completely the range of conformations visited by the TNF1 peptide.)

### A hypothetical model of peptide-protein binding

It should be clear from the evidence reviewed in this chapter that peptide binding to proteins is far too complex a process to be captured accurately by a simple system of differential equations based on the usual combination of mass action and heuristically determined parameters. Assuming hypothetically the tentative conclusions expressed in the preceding discussion, an improved model should address at least the following issues:

1. *Peptide conformation*. The free peptide in solution inhabits a conformational ensemble of indeterminate diversity. The fraction of these conformations capable of leading to a stable complex is likely very small and difficult to estimate, making the TΔS component of ΔG a matter of guesswork.

2. *Rotation and translation*. Even if the conformations are compatible, a collision will not result in even transient complex formation unless both the peptide and the protein are positioned and oriented such that contact occurs between regions of the peptide and loci on the protein that are capable of interacting such that the aggregate ΔH of all moieties in contact is sufficiently negative to overcome the entropic penalty.

3. *Multiplicity of binding configurations*. A single binding site cannot be assumed. There are potentially multiple ways of choosing a conformation of the peptide and a region on the protein target that can be placed in (perhaps partial) contact such that a negative ΔG results.

4. *Post-encounter transitions*. It must be expected that a peptide, once in at least transiently stable contact with the protein, may undergo changes in position and/or conformation to find deeper energy minima to the extent that the energy required to cross any intervening barriers can be supplied by thermal fluctuations.

5. *Distributed affinity*. Consider a "snapshot" of a single bound peptide-protein complex: the affinity is determined (ignoring entropic effects) by ΔH, which may be regarded as the sum of the ΔH contributions of all of the various contacts. As will be shown in the next chapter, these contributions are far from uniform, and it is common for a few "hot spot" residues, often separated by stretches of relatively low affinity, to account for most of the aggregate ΔH.

6. *Steady state dynamics*. As will be shown in greater detail in the next chapter, peptides, even when stably bound, are far from immobile. A peptide side chain moiety in a non-bonded interaction with the protein target can often, by a minor shift requiring energy well within the bounds of thermal fluctuations, find and make another non-bonded interaction with no significant net loss of energy.

To these might be added further complicating factors such as flexibility of the protein at scales ranging from minor side chain positional adjustments to large movements around hinge regions; the potential for the peptide to bind to the modified protein surface created when another molecule of the same peptide has already bound; and coordination by ions present in the solution.

The main conclusion to be drawn is that a model that represents peptide-protein binding in terms of a single paired configuration is inadequate; a probabilistic model is required. The following thought experiment will serve to illustrate the concept: Consider a conformational ensemble of free peptide in solution, from which a single conformation is selected at random. Choose also a random selection from the conformational ensemble of the protein target. Choose random orientations in space for both the peptide and the protein, choose random vectors of approach with random translations between them, and bring the two into contact along those vectors at speeds randomly chosen based on a Boltzmann distribution.  Now allow the complex to relax to its local energy minimum (for example, by molecular dynamics starting with the initial contact state and velocities), and determine the ΔG corresponding to that state.

If the foregoing process could be repeated a sufficient number of times to sample the problem space adequately, one could in principle construct an energy landscape expressing average ΔG as a function of the bond rotations in the peptide and the relative rotational and translational initial contact positions of peptide and protein.  The working hypothesis underlying the studies presented in the chapters to follow is that such a landscape would typically have many more than a single local energy minimum, many more than one of those minima would have negative ΔG, and relatively low-barrier transitions may exist between nearby minima. The result would be an ensemble of bound complexes corresponding to the various energy minima, with abundances presumably approximating a Boltzmann distribution.

It is not possible to perform the experiment just described on 20-residue peptides using currently available computing resources. It is, however, entirely feasible to perform an analogous experiment by fragment based methods, in which energy landscapes along the lines described are computed for overlapping trimer fragments and combined to produce a composite spatial mapping of binding probabilities onto the protein surface. As will be seen in Chapter 4, the resulting maps predict the predominant binding sites of peptides in x-ray structures of peptide-protein complexes with an accuracy superior to that of other reported methods. Predictions using the same probability mapping method are shown in Chapter 5 to be consistent with binding site measurements made (by others) by cross-linking a CIM-10K library peptide to AKT-1 protein and determining the surface residues involved by mass spectrometry.

Before describing the energy mapping experiments, however, Chapter 3 will further develop the theoretical foundations of the proposed model, by examining in detail the actual measurable characteristics of peptide-protein interfaces for which solved structures are available.

# CHAPTER 3: CHARACTERISTICS OF PEPTIDE-PROTEIN INTERFACES

## Abstract

PDB structures involving peptides bound to proteins furnish a rich source of information about the characteristics that contribute to stable peptide-protein associations, information that is potentially useful in engineering peptide ligands.

A dataset ("PPRMint", for "Peptide-Protein Reduced Minimized Interfaces") of 3,924 energy-minimized interfaces was assembled from the PDB, each interface comprising a single peptide chain of at least 8 and not more than 32 residues, together with the segments of the protein chain(s) with which each is associated. These were analyzed to extract descriptors of geometry, interactions, and other characteristics potentially affecting affinity. An energy model was trained using 75 interfaces for which experimental affinity data was available from the PDBBind database, and the model was used to estimate energy contributions by residue and interaction type. The PPRMint dataset represents all peptide-protein interfaces in the PDB as of July 27, 2008 that met the relevant selection criteria. It has been made available in a public web repository, together with a relational database, readily accessible using Microsoft Office, containing the detailed results of an analysis of each interface, peptide residue, and non-bonded interaction. Based on an analysis of the peptide-protein interfaces in the dataset, peptide and interface characteristics are identified and summarized that

appear to influence affinity and may provide useful guidance in the design of peptide ligand libraries and the selection of leads for optimization.

The results presented here suggest that peptide ligands should target protein surface sites that are rich in aromatic and charged residues and that preferably lie in concavities, and that the peptides themselves should also be rich in aromatic and charged residues, and of intermediate flexibility. Optimum attainable affinities appear to be in the 10 nM range, more or less independent of length for peptides in the range from about 12 to 20 residues.

### Background

Despite considerable and growing interest in peptides for therapeutics and diagnostics, discovery of peptide ligands remains much more a matter of trial and error than of intelligent design. One potential source of insight into the mechanisms of peptide binding is the large and growing pool of PDB structures that contain peptide-protein interfaces. Although a few recent studies have begun mining this vein[185-187], and although there is an extensive literature describing the properties of protein-protein interfaces [188-197], protein interfaces with intrinsically disordered proteins [70, 71, 198-200], interactions among the residues of folded proteins [201], secondary structure characteristics of folded proteins [202], and other similar observations derived from analyses of PDB-derived datasets, many of the characteristics of peptide-protein interfaces have not yet been explored, and the factors that influence affinity and specificity remain a matter of debate. In the present work, a dataset ("PPRMint", for "Peptide-Protein

Reduced Minimized Interfaces") of 3,924 protein interfaces with 8- to 32-mer peptides was extracted from PDB structures and analyzed. Descriptors were computed relating to the geometry of the interfaces, the geometry and energetics of the non-bonded interactions between peptide and protein, residue frequencies and spatial relationships between peptide and protein residues, estimated binding energy contributions at the interface and residue levels, and other factors. This information, together with other descriptive information, has been recorded in the tables of a Microsoft Access relational database which has been made available for download, together with the structures of the 3,924 extracted and minimized interfaces in a standardized PDB-compliant format that facilitates further analysis (see Appendix 1). This data has been used to inform exploration for heuristics, presented here, that might be used to guide the discovery of peptide ligands.

For clarity regarding terminology: Each of the interfaces in the PPRMint dataset consists of a contiguous peptide chain, uniformly designated the "P" chain in the extracted interface files, and a second chain, designated the "I" (interacting) chain that is a composite of all residues that are within a 25Ǻ distance cutoff from any part of the peptide chain and that belong to the molecular entity to which the peptide is bound. The "I" chain may contain residues derived from more than one chain in the original PDB structure, and gaps due to the exclusion of residues beyond the cutoff distance may be present. The subset of I chain residues that are actually in close contact with the peptide (having any atom within 4Ǻ of any atom of the peptide) are referred to collectively as the "I-site"; in effect, the "I-site" residues form the surface patch to which the peptide is bound. (When a peptide residue is described as ‘in contact’ with an I-site residue, it is meant that at least one atom in the peptide residue is within 4Ǻ of an atom of the I-site residue.) The entire protein to which the peptide is bound is referred to as the “target”.  The problem of redundancy was handled by clustering structures into groups each comprising one or more highly similar structures, rather than by restricting the dataset to a single exemplar of each structure; these are referred to as “redundancy groups”. Finally, where appropriate, when referring to values or properties for which entries exist in the tables of the database, the name of the relevant database field is shown in parentheses.

Energy estimates were made using two distinct models.  A computer program,  “PopTop” (for “POlyPeptideTOPology”), was written that analyzes a PDB structure,  identifies and quantifies each of the various individual interactions that might be expected to contribute to binding, and computes various descriptors relating to size, compactness, and surface area. Source code for PopTop is made available at http://www.innovationsinmedicine.org/pprmint/source/PopTop. The “PopTop model” refers to a 9-parameter energy model employing descriptors computed by the PopTop software, with weights fitted to affinity data using a training set of 75 peptide-protein interactions from the PDBBind database of Wang et al. [203, 204].  To provide a basis for comparison, a second model, the “Autodock-based model”, was employed, using energy components obtained using the Autodock molecular docking application [205, 206], in combination with size-related descriptors obtained from PopTop, with weights trained using the same training set.

### Methods

A comprehensive dataset of PDB structures was assembled containing peptides in the 8 to 32 residue size range bound to protein surfaces.  From these were extracted the peptide-protein interfaces, which were energy-minimized.  On the resulting 3,924 interfaces, containing 61,130 peptide residues, geometric measurements were made and other descriptors were extracted, interaction energies were predicted using two distinct models, and the resulting data was uploaded to a Microsoft Access relational database to facilitate inquiry concerning various interface properties.

### Assembly of PPRMint dataset

The 52,103 files contained in the Protein Data Bank as of July 27, 2008 were downloaded in PDB format and pre-screened to identify those possibly containing peptide chains interacting with other peptides or proteins.  After excluding the 56 files lacking coordinate data, the remaining 52,047 files were found to contain 128,079 chains, of which 11,413 chains in 6,084 files had lengths in the range 8 to 32 residues (based on the number of residues shown in the corresponding SEQRES lines [207]), inclusive.  Of these 6,084 files, those in which the peptide chain was the only chain present (1,240 files) were excluded; also excluded were those in which all of the chains in the 8 to 32 length range had at least one residue that was not one of the standard 20 amino acids (3,570 files, many containing nucleotide chains, and others whose peptide chains contained modified or non-natural residues), and those in which the only possible interacting chain(s) were either shorter than 8 residues or contained no standard amino acid residues (24 files). (A number of files belonged to more than one of the foregoing excluded categories.) Following this screen, 1,813 files remained. These files were then hand-screened to categorize by interaction type and to examine each structure so as to exclude any non-conforming interfaces missed by the automated pre-screen. 1,351 files remained after excluding 170 files involving peptides complexed to other peptides, 117 files involving insulin complexes, 126 files in which peptides were wholly or partially buried in the protein, 60 files involving peptides as part of ribosomal or other large complexes, and 103 files rejected for other reasons (with some files again being rejected on more than one criterion).

### Extraction and minimization of interfaces

From the 1,351 PDB files, interfaces were extracted by identifying, in each file, all peptide chains having a length of at least 8 and not more than 32 residues, and writing the ATOM lines corresponding to each such peptide to a separate file together with all ATOM lines corresponding to all residues having any atom within 25Å of any atom in the peptide. The 25Å cutoff distance was chosen so as to reduce the size and complexity of the interface files without excluding parts of the structure close enough to affect significantly the electrostatic terms of the energy models. For uniformity, in the interface files, the peptide chain is always given a chain ID of “P”, and the ATOM lines belonging to the protein portion of the interface are given a chain ID of “I”. This may result in the “I” chain containing segments derived from more than one chain in the original structure, and makes it necessary for subsequent analyses to correctly handle backbone gaps and to avoid modifying charges on gap-boundary-terminal amines and carboxyls.

Because many PDB structures involve crystallization arrangements in which multiple complexes are stacked closer than 25Å apart, it was then necessary to visually inspect and correct all interfaces in which the “I” chain contained segments from more than one chain in the original PDB file, to remove any nearby segments belonging to adjacent structures that are not part of the specific object with whose surface the peptide is interacting. After excluding interfaces in which the interacting surface was not clearly determinable or in which other data integrity issues were found, a total of 3,924 corrected interfaces were extracted from the 1,351 PDB files. These have been made publicly available at http://www.innovationsinmedicine.org/pprmint/dataset. Naming conventions and file contents are described in the file readme.pdf in that directory.

When energy prediction models were applied to a subset of the interface files corresponding to peptide-protein interactions for which published affinity data is available (as described below), it was found that the Autodock energy function produced poor results, traceable to improbably high positive van der Waals energies for a few atom pairs. Molecular dynamics energy minimization of the interfaces was tested under several choices of parameters, as was modification of the energy function to impose an arbitrary cutoff on positive atom pair energies. Minimization produced the best agreement with published values, and was therefore performed on all 3,924 interfaces, using Amber 9 [208]. Each interface was prepared and checked using tleap and any errors corrected, followed by 250 cycles of steepest descent and 250 cycles of conjugate gradient minimization with cut = 12.0Ǻ.

### Construction of database

A relational database was constructed containing an 'Interfaces' table having one record for each of the 3,924 interfaces; a 'Residues' table having one record for each of the 61,130 residues belonging to the peptides in the 3,924 interfaces, with each keyed to the corresponding interface record; and 'Hbonds', 'SaltBridges', and 'CationPi' tables having one record for each of the 30,583 interactions satisfying the gross screening criteria for hydrogen bonds, the 7,090 interactions satisfying the criteria for salt bridges, and the 11,387 interactions satisfying the criteria for cation-pi interactions (5,078 distinct interactions discounting furcations), respectively, each keyed to the records of the participating residues and including the details of the bond geometry of each. Included in this database are the results of energy estimates and other computations described here. The database was constructed using Microsoft Access, so as to make it readily accessible both to end users needing the simple, spreadsheet-like behavior available to any Microsoft Office user, and to those wishing programmatic access using an ODBC connection. The database, together with descriptive material, has been made publicly available at http://www.innovationsinmedicine.org/pprmint/database. (As used herein, the “PPRMint dataset” refers to the 3,924 PDB format interface structure files; the “PPRMint database” refers to the relational database.)

### Extraction of descriptive data and energy components using PopTop

Using the PopTop software, each of the 3,924 interfaces was analyzed and a number of descriptive quantities were computed, both at the interface level and the residue level, of which a weighted sum of the most informative are incorporated into an overall energy scoring function, with weights calibrated by fitting to published affinity data. PopTop identifies and reports on the various interactions that may contribute energy by examining all pairings of P chain atoms with I chain atoms where the distance between the two is less than 5Ǻ plus the sum of the van der Waals radii of the two. For all such pairings, an evaluation is made to determine whether a hydrogen bond, salt bridge, or pi-cation interaction is present and if so to estimate its energy, and also to estimate the energy corresponding to the hydrophobic, electrostatic, and van der Waals forces between the pair.  PopTop also computes and reports other quantities describing geometric properties, such as peptide end-to-end length, peptide radius of gyration, and difference in solvent accessible surface area between the bound complex and each of the interacting entities taken separately.

*Interaction-based descriptors*

Any pairings of specific atoms in specific residues whose identities conform to the atlas of main chain and side chain hydrogen bond types as described in [209] are evaluated according to the hydrogen bond energy model described in [210, 211], which determines interaction energy as a function of geometry and hybridization:

$$E_{HB} = V_0 \left\{ 5\left(\frac{R_0}{R}\right)^{12} - 6\left(\frac{R_0}{R}\right)^{10} \right\} F(\theta, \phi, \varphi)$$

where $V_0$ = 8 kcal/mol and $R_0$ = 2.80Å, R is the distance between the hydrogen and the acceptor, θ is the donor-hydrogen-acceptor angle, φ is the angle made by the hydrogen, the acceptor, and the atom to which the acceptor is bonded, ψ is the dihedral angle formed by donor, hydrogen, acceptor, and the atom to which the acceptor is bonded, and *F(θ,φ,ψ)* depends upon donor and acceptor hybridization as follows:

$sp^3$ donor, $sp^3$ acceptor: $F = \cos^2\theta \exp(-(\pi-\theta)^6)\cos^2(\phi - 109.5)$

$sp^3$ donor, $sp^2$ acceptor: $F = \cos^2\theta \exp(-(\pi-\theta)^6)\cos^2\phi$

$sp^2$ donor, $sp^3$ acceptor: $F = \cos^4\theta \exp(-2(\pi-\theta)^6)$

$sp^2$ donor, $sp^2$ acceptor: $F = \cos^2\theta \exp(-(\pi-\theta)^6)\cos^2(\max[\phi, \varphi])$

Any pairings of lysine ζ-N, arginine η-N, or an N-terminal backbone N atom with an aspartate δ-O, glutamate ε-O, or a C-terminal carboxyl O atom between the peptide chain and the I-site at a distance of 5Å or less are evaluated according to the model described in [211]:

$$E_{SB} = V_0 \left\{ 5\left( \frac{R_0}{R + 0.375} \right)^{12} - 6\left( \frac{R_0}{R + 0.375} \right)^{10} \right\}$$

where $V_0$ = 8 kcal/mol, $R_0$ = 3.2Å, and R is the distance between cation and anion.

Both energy functions are in the form of 12-10 Lennard-Jones potentials, and therefore reach high positive energies quickly as distances become closer than optimal. Since it is hypothesized that peptide interfaces are flexible enough that any such overly close pairings would be eliminated by movement, and given the inaccuracies inherent in PDB coordinates, it was regarded as unrealistic to impose large positive energy penalties merely because of slightly smaller than ideal spacing, so the foregoing models were modified to report energies based on the optimum distance where the measured distance is closer than optimum.

Since the overall energy model is calibrated by fitting to training data, since the relative predictive accuracy of various theoretical energy models is in any case a matter of debate, and since PDB coordinates are at best only approximately accurate, the choice of energy models for estimating hydrogen bond, salt bridge, and cation-pi energies was motivated mainly by a desire for a qualitatively realistic relationship between energy and optimality of the interaction geometry, without necessarily expecting absolute quantitative accuracy. PopTop identifies and evaluates potential cation-pi interactions between the aromatic rings of phenylalanine, tyrosine, and tryptophan and the charged amines of lysine, arginine, and the backbone N terminus, for all pairings where the distance between cation and ring centroid is ≤ 8Å. (Pairings involving histidine were not evaluated, since its participation in cation-pi interactions depends upon its protonation state, see [212], which is not reliably determinable from PDB data.) For the cation-pi energy function, a heuristic expression was devised in the general form of a 5-4 Lennard-Jones potential that approximates the theoretical relationship between energy and cation position relative to the aromatic ring for a benzene – $NH_4^+$ system as reported graphically in Figure 2(d) of [213]:

$$E_{CP} = 16.5\left\{5\left(\frac{R_0}{R_{adj}}\right)^5 - 6\left(\frac{R_0}{R_{adj}}\right)^4\right\}\left(\frac{1}{\left(1+\left(\frac{\theta}{50}\right)^{2.6}\right)}\right)$$

where $R_{adj} = R - \left(\frac{\theta}{50}\right)^{1.4}$, R is the distance between the centroid of the aromatic ring and the cationic atom, θ is the angle between a normal to the plane of the ring and a vector from the centroid of the ring to the cationic atom, and $R_0$ is 2.8Å.

PopTop employs a hydrophobicity / desolvation energy model that is based generally on the approach described by Fernandez-Recio et al. [214], in which the change in desolvation energy on binding is given by:

$$E_{desolv} = -\sum_i \sigma_i ASA_i$$

where $ASA_i$ is the change in accessible surface area (ASA) of atom *i* on binding, $\sigma_i$ is an ‘atomic solvation parameter’ (ASP) measuring the contribution to solvation energy per unit ASA for atoms of the type of atom *i* as derived from octanol/water transfer energies of N-acetyl amino acid derivatives, and the summation is over all atoms whose ASAs change on binding. Because one of the goals of the project was to obtain insights about how the various residue types in the peptide participate in binding, making it necessary to estimate energy contributions at the residue level as well as for the entire interface, the hydrophobicity model was modified to address two main issues. First, to determine the overall energetic favorability of the binding of an individual residue of the peptide, it is necessary to also take into account the desolvation of the protein residue(s) to which it binds, and atoms belonging to the peptide residues may often affect the desolvation of atoms belonging to more than one protein residue. Second, the $\sum_i \sigma_i ASA_i$ model implicitly assumes that every atom is transferred from an aqueous environment to an octanol-like hydrophobic environment, which is not necessarily the case. As others have noted [54, 215], it may therefore be preferable to take into account the hydrophobic complementarity of each atom with the microenvironment in which it finds itself after binding. (A further reservation is that the computation of ASA on the unbound structures assumes that their conformations in solution would be the same as the bound conformation, obviously unlikely to be true on average for the peptide and perhaps only approximately true for the protein. Attempting to predict the distribution of peptide conformations in solution is beyond the scope of the current project. The problem might be addressed in part by generating distributions of conformations of each peptide in solution by molecular dynamics, as described in Chapter 2, but doing so for 3,924 peptides was deemed impracticable.)

PopTop computes and reports two distinct estimates of energy relating to hydrophobic interactions. The first measure (EDS) uses the $\sum_i \sigma_i ASA_i$ model to compute the desolvation energy of the atoms and residues of the peptide, and also reports, for each peptide residue, an estimate of the desolvation energy of that part of the protein chain allocable to the peptide residue. This allocation is made by examining each pair of interacting atoms and computing the proportion of the ASA change of each attributable to the other. It is assumed that given two interacting atoms $a_1$ and $a_2$ separated by a distance *d*, the surface area of $a_1$ from which water is excluded by $a_2$ can be estimated by computing the area of the spherical cap of radius $r_1$ equal to the van der Waals radius of atom $a_1$, whose boundary is defined by the point of contact of a 1.5A probe sphere contacting both atoms. This measure is not entirely accurate, even leaving aside the usual reservations about treating atoms as rigid spheres, because the spherical cap on $a_1$ from which water is excluded by $a_2$ may include area from which water would also be excluded by another nearby atom $a_3$; however, it does provide a reasonable basis on which the total desolvation energy contribution of the protein side of the interface can be prorated among the peptide atoms. PopTop reports this “opposite chain” allocable desolvation energy (ocEDS) for each peptide residue and for the entire peptide.

Since the foregoing analysis provides an estimate of the amount of surface area of each peptide atom affected by each nearby protein atom, it is then possible to estimate the extent to which each atom is positioned next to other atoms having compatible hydrophobicity properties, and adjust energy estimates accordingly. For each pairing of atom $a_1$ with an interacting atom $a_2$ in close enough proximity to produce an ASA change on binding, the EDS value calculated for $a_1$ (the product of $a_1$’s ASA change attributable to $a_2$ and $a_1$’s ASP $\sigma_1$) is taken as the estimate of the maximum desolvation energy change that would occur if $a_2$ provided an octanol-like hydrophobic environment, and then multiplied by an additional factor $\mu = 1 - (1.8181\ \varphi)$, where $\varphi$ is the hydrophilicity index of atom $a_2$ adapted from the hydrophilicity scale of Kuhn et al. [216]. The values of $\varphi$ for various atom types range from a value of 0 for hydrophobic atoms such as the backbone carbons of the very hydrophobic residues isoleucine and leucine, to a maximum of 0.635 for the η oxygen of tyrosine. Where atom $a_1$ is positioned next to a very hydrophobic atom having $\varphi = 0$, then $\mu = 1$, and the resulting adjusted desolvation energy (EHYB) is equal to the computed maximum desolvation energy EDS.  An assumption is made that the γ oxygens of serine ($\varphi = 0.491$) and threonine ($\varphi = 0.601$) have approximately water-like properties, so at their average $\varphi$ of 0.55, the foregoing heuristic expression yields $\mu = 0$, and the EDS attributable to this pairing is disregarded, since $a_1$ has in effect moved from an aqueous environment in the solvent to another aqueous-like environment in the interface. μ becomes slightly negative when $a_1$ is adjacent to atoms having φ > 0.55, reflecting a transition from aqueous solvent to an even more hydrophilic environment. PopTop reports both the EHYB of each peptide residue and also each peptide residue's allocable share (ocEHYB) of the energy change attributable to protein residues to which it is adjacent.

PopTop's electrostatic energy model follows the electrostatic component of the Autodock force field [205]. Charges are assigned to each atom according to its residue and type based on the Amber ff94 force field parameters [217]. Each atom $a_i$ in the peptide is assigned an electrostatic energy equal to the sum of the energies computed for each pairing of $a_i$ with each atom $a_j$ of the protein within a distance cutoff of approximately 8Å (5Å plus the sum of van der Waals radii), according to the Coulomb relation:

$$EES_{ij} = k\frac{q_i q_j}{\varepsilon\left(r_{ij}\right)r_{ij}}$$

where $k$ is the Coulomb constant, $q_i$ and $q_j$ are the charges on $a_i$ and $a_j$ respectively, $r_{ij}$ is the distance between $a_i$ and $a_j$, and $\varepsilon(r_{ij})$ is a distance-dependent dielectric constant as described in [218, 219]. $\varepsilon(r_{ij})$ is computed as

$$\varepsilon\left(r\right) = A + \frac{B}{1 + ke^{-\lambda Br}}$$

where B = $\varepsilon_0$, the dielectric constant of bulk water (78.4 at 25°C), A = -8.5525, k = 7.7839, and λ = 0.003627 $Å^{-1}$.

PopTop computes van der Waals energies for each pair of interacting atoms $a_i$ and $a_j$ using a Lennard-Jones 12-6 potential closely patterned after the van der Waals component of the Autodock force field model [205]:

$$EVDW_{ij} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}$$

$A_{ij}$ and $B_{ij}$ are parameters derived from the Amber force field [220] each of whose magnitude depends on the identities of atoms $a_i$ and $a_j$ as H, C, O, N, or S:

$$B_{ij} = -2\varepsilon_{ij} r_{\min}^{6}$$

$$A_{ij} = 0.5 B_{ij} r_{\min}^{6}$$

$$\varepsilon_{ij} = -W_{VDW}\sqrt{\varepsilon(i)\varepsilon(j)}$$

$$r_{\min} = \frac{r_{\min}(i) + r_{\min}(j)}{2}$$

where $r_{min}(i)$ and $\varepsilon(i)$ depend on the identity of atom $a_i$ as H, C, O, N, or S as follows: If $a_i$ is H, $r_{min}(i) = 2.0$, $\varepsilon(i) = 0.02$; if $a_i$ is C, $r_{min}(i) = 4.0$, $\varepsilon(i) = 0.15$; if $a_i$ is O, $r_{min}(i) = 3.2$, $\varepsilon(i) = 0.2$; if $a_i$ is N, $r_{min}(i) = 3.5$, $\varepsilon(i) = 0.16$; and if $a_i$ is S, $r_{min}(i) = 4.00$, $\varepsilon(i) = 0.2$. $W_{VDW}$ is an Autodock weight factor equal to 0.1662. (Weight factors on individual energy components are arbitrary since, as described below, new weights are computed by training the overall model specifically to peptide-protein binding data.)

### *Geometric descriptors*

PopTop also computes and reports several geometric measures that were hypothesized to be informative for purposes of energy prediction. These include:

1. Plen: the length of the peptide, expressed as the number of residues in the chain for which coordinates are present in the PDB file (note that in peptide-protein interfaces it is not uncommon for one or both ends of the peptide to be unbound and unstructured, with no coordinates given in the PDB structure).
2. SRlen: the length of the peptide, expressed as the number of residues shown in the SEQRES lines of the original PDB file, which should ordinarily reflect the full length of the peptide used in the crystallization or NMR experiment [221].
3. Rg: the radius of gyration of the peptide chain (necessarily based on the segment for which coordinates are given), computed as the square root of the mean of the squared distances between each $C_\alpha$ atom and the centroid of all $C_\alpha$ atoms.
4. Endlen: the peptide chain end-to-end distance, computed as the distance between the $C_\alpha$ atoms of the N-terminal and C-terminal residues for which coordinates are present in the PDB file.
5. ASA, ASAunb, and ocdASA: respectively, the ASA of the peptide chain in its bound form; the ASA of the peptide chain taken by itself, without the interacting chain(s) present; and the total difference in ASA of the protein with and without the peptide present.

### Energy descriptors using Autodock

A Python script was written to set up inputs for each interface, run the Autodock docking software based on the bound position of the peptide, and parse energy values from the Autodock output. Unfortunately, the most detailed Autodock output (written by Autodock to a *.dlg file) provides only three independent descriptors: vdW_HB_DS_en, which comprises the energy attributable to van der Waals forces, hydrogen bonds, and desolvation (all expressed in a single quantity); es_en, representing energy due to electrostatic forces; and tors_en, the torsional free energy, essentially a fixed multiple of the number of rotatable bonds in the ligand (intended to account for the entropic penalty for restricting bond rotations as a result of binding). Autodock outputs the first two on an atom by atom basis, allowing values of each to be parsed out and allocated to each residue of the peptide.

### Fitting of energy scoring functions to peptide-protein affinity data

A dataset for use in training the weights of both the PopTop and Autodock energy scoring functions was obtained by extracting affinity data for the relevant interfaces from the PDBBind database [203, 204]. The entire PDBBind database was downloaded as of October 23, 2008, consisting of more than 3000 interactions (most pertaining to small molecule and other interactions not involving peptides), and data was extracted for the 75 interactions whose PDB ID's corresponded to peptide-protein interfaces present in the PPRMint dataset. Because of multiple models in NMR structures and because many crystal structures contain multiple instances of the same peptide-protein pairing, 250 interfaces in the PPRMint dataset matched peptide-protein pairings in the PDBBind 75-peptide training set. Descriptors were extracted separately for each of these 250 interfaces, and were averaged in cases of multiple interfaces representing the same PDBBind interaction. The stated $pK_D$ values (converted to energies in kcal/mol) from the PDBBind database were used to train the energy models.

Of the descriptors reported by PopTop as described above, nine were selected as providing the most informative fit to training data. Six of these were measures intended to reflect the energy contribution of hydrophobic interactions (EHYB and ocEHYB), hydrogen bonds (EHB), salt bridges (ESB), cation-pi interactions (ECP), and van der Waals forces (EVDW) as described above, and three, radius of gyration of the peptide (Rg), length of the peptide expressed as number of residues (Plen), and solvent accessible surface area of the peptide in its bound state (ASA), relate to the dimensions of the peptide and the interface. The predicted binding energy is given by

$$\Delta E = c + \sum_i w_i d_i$$

where $c$ is the constant basis term, $d_i$ are the nine descriptors, and $w_i$ are the corresponding weights which were trained by regression fitting to the energies derived from the PDBBind database for the training set.

A similar linear model was evaluated using six descriptors: the three obtained using Autodock (vdW_HB_DS_en, es_en, and tors_en) together with the same three geometric descriptors as used for the PopTop energy model (Rg, Plen, and ASA), trained similarly using the PDBBind training set. Regression statistics were computed for each of the two scoring functions, as well as several descriptor subsets (see Table 1), for the training set as a whole and for leave-one-out cross-validation. The weights thus determined for the PopTop and the Autodock-based energy functions were then applied to the corresponding descriptors over the entire 3,924-interface PPRMint dataset and total energies (PT_Ettl and AD_Ettl, respectively) computed for all interfaces and for all peptide residues in each interface. In computing the residue-level total energies, the energy contributions due to descriptors that relate to the entire interface as a whole rather than to individual residues (such as radius of gyration, peptide length, ASA, and the Autodock torsion term) were allocated equally over all peptide residues that are in contact with the protein.

| | **Training Set** | | | **Cross-validation** | | |
|---|---|---|---|---|---|---|
| **Energy Function** | **$r_P$** | **ME** | **SE** | **$r_P$** | **ME** | **SE** |
| PopTop, 9-parameter | 0.72 | 1.07 | 1.37 | 0.60 | 1.24 | 1.60 |
| AutoDock, 6-parameter | 0.66 | 1.21 | 1.50 | 0.54 | 1.36 | 1.68 |
| PopTop, 6 energy descriptors | 0.48 | 1.42 | 1.73 | 0.32 | 1.57 | 1.90 |
| AutoDock, 3 energy descriptors | 0.35 | 1.50 | 1.85 | 0.13 | 1.60 | 2.04 |
| 3 size descriptors only | 0.61 | 1.28 | 1.57 | 0.55 | 1.36 | 1.66 |

Table 1. Energy function regression statistics. Pearson correlation coefficient (rP), mean absolute error (ME) (kcal/mol), and standard error (kcal/mol) for the energy functions underlying the energy statistics reported herein (Poptop 9-parameter and Autodock).

### Assessing redundancy and clustering of related structures

For each interface, an interface ‘signature’ was computed consisting of a colon-delimited text string in which each field begins with the single-letter amino acid code of the peptide residue and is followed by the amino acid codes of all protein residues within 4Ǻ of the peptide residue, in descending order by distance between the two closest atoms in the residue pair). Each interface was assigned to a peptide redundancy group by assembling an n × n similarity matrix M (where n is the total number of interfaces in the dataset), each of whose elements $m_{ij}$ is set to 1 if the signature of interface i is similar to that of interface j, and to 0 otherwise. Two signatures are automatically deemed similar if the two interfaces are alternate models of the same NMR structure. Otherwise, an ungapped alignment is performed, in which the first sequence is compared to the second in each possible alignment having an overlap of at least 5 residues; if in the best alignment there are any mismatches in the overlap region, the two signatures are deemed dissimilar. Each interface is assigned a peptide redundancy group ID number (‘Qgrp’), with groups determined from the connectivity matrix by assigning to a redundancy group all interfaces corresponding to the elements of the row of M having the highest 1-norm, setting to zero all rows and columns of M corresponding to the interfaces so selected, and repeating these steps until no non-zero elements remain. Each interface is also assigned a second peptide redundancy group ID number (‘Pgrp’) in an identical fashion except that the similarity criteria are made more stringent by regarding as a mismatch not only any residue mismatch in the overlap region but also the presence of any surface-bound residue outside the overlap region.

The Pgrp and Qgrp groupings are focused on similarity of the peptide sequences. A further grouping (‘Rgrp) is assigned, using the same clustering strategy, based on similarity of the overall interface as described by the signature. For two interfaces to be deemed similar for this grouping, both must belong to the same Pgrp and, in addition, the set of protein residues to which each peptide residue is bound must agree in both interfaces for at least half the bound residues in the peptide.  The set of protein residues to which two peptide residues are bound are deemed to agree if and only if the closest protein residue to each is among the closest three of the other (to allow for minor variations in ordering of interacting residues by distance).

The 61,130 individual peptide residues were also clustered into 14,592 residue-level redundancy groups (‘Rgrp’ in the Residues table) according to the following criteria: Two peptide residues are assigned to the same redundancy group if and only if (1) their parent interfaces belong to the same interface level redundancy group (Rgrp), (2) the number of I-site residues interacting with each (within 4Ǻ) differs by no more than two, (3) each residue type interacting with the peptide residue having the fewer interacting residues is also present in the set of residues interacting with the peptide residue having the more interacting residues, and (4) the flanking residue types in both directions are the same for both residues.

### Computation of isoelectric points and charges

Net charges on peptides (Pchg) were computed based on the full peptide sequence as shown in the SEQRES lines of the PDB file at pH 7.4 using:

$$Q = \sum_r \frac{q_r n_r}{1 + 10^{q_r (pH - pK_r)}}$$

where *r* refers to the moieties capable of protonation, $q_r$ is -1 for residues D, E, C, and Y and the C-terminal carbonyl, and +1 for residues H, K, R and the N-terminal amine, $n_r$ is the number of times moiety *r* appears in the sequence, and $pK_r$ is the pK value of moiety *r* as given in [222]. Net charges on I-sites (Isitechg) were computed similarly for the set of residues comprising the I-site, assuming no charged N- or C-termini. The pI of each peptide was computed by performing a binary search for a pH value giving a charge *Q* in the range (-0.001, 0.001) as computed by the foregoing equation.

## Results

### Energy scoring function regression fitting and validation

As shown in Table 1, an energy model based solely on the 6 PopTop descriptors relating to estimated energies of non-bonded interactions, or on the three aggregated parameters output by Autodock alone, provided a Pearson correlation ($r_P$) with training data of 0.48 and 0.35, respectively and mean absolute error of 1.32 and 1.50 kcal/mol, respectively). With three descriptors of the peptide's size and geometry (radius of gyration (Rg), length (number of residues) (Plen), and solvent accessible surface area of the peptide (ASA)) added to both models, results were $r_P = 0.72$ and ME = 1.07 kcal/mol for the PopTop model and $r_P = 0.60$ and ME = 1.24 for the Autodock-based model. On leave-one-out cross-validation, the 9-parameter PopTop model produced a correlation of 0.60 and a mean absolute error of 1.24 kcal/mol (0.90 $pK_D$ units); the 6-parameter Autodock-based model produced a correlation of 0.54 and a mean absolute error of 1.36 kcal/mol (0.99 $pK_D$ units). Table 2 shows the fitted descriptor weights for both models, as well as the relative contribution of each descriptor to the computed total interface energies. The correlation between the PopTop and Autodock-based models was high ($r_P = 0.87$ over the 75 interfaces of the PDBBind-based training set, $r_P = 0.88$ over the 3,924 interfaces of the full dataset).

### General characteristics of PPRMint dataset

The 3,924 interfaces comprising the PPRMint dataset make up 941 distinct redundancy groups by the most restrictive measure of similarity used (Rgrp); these encompass a broad cross-section of target proteins, which may be loosely categorized by type as shown in Figure 10. Immunological molecules (antibodies and MHC's) account for about a quarter of the dataset, enzymes and receptor proteins together account for about another quarter, conserved domains (PDZ, SH2, and SH3) account for a few percent, and the remaining ~40% encompass a range of other protein types. (The type categories in the database were obtained in a partially automated fashion from PDB header descriptions, and are intended to provide a general breakdown; an item by item verification has not been undertaken.)

| | | **Contribution to interface total energy (kcal/mol)** | | |
|---|---|---|---|---|
| | **Weight** | **Mean** | **Range** | **σ** |
| **PopTop 9-Parameter Energy Function** | | | | |
| Hydrophobic interaction energy: attributable to peptide (EHYB) | -0.0934 | -0.10 | 1.06 | 0.23 |
| attributable to protein (ocEHYB) | -0.2176 | -0.60 | 1.82 | 0.45 |
| Energy due to hydrogen bonds (EHB) | -0.0330 | 0.83 | 2.32 | 0.52 |
| Energy due to salt bridges (ESB) | 0.0345 | -0.38 | 1.47 | 0.29 |
| Energy due to cation-pi interactons (ECP) | 0.0252 | -0.41 | 1.44 | 0.39 |
| Energy due to van der Waals forces (EVDW) | 0.0755 | -1.22 | 1.38 | 0.42 |
| Radius of gyration (Rg) | 0.5445 | 3.89 | 3.64 | 0.74 |
| Peptide accessible surface area (ASA) | 0.0038 | 2.31 | 3.51 | 0.87 |
| Length of peptide (no. of residues) (Plen) | -0.4938 | -4.89 | 6.42 | 1.20 |
| Constant basis term | -7.8259 | -7.83 | 0.00 | 0.00 |
| **Autodock-based 6-parameter Energy Function** | | | | |
| Energy due to van der Waals forces, hydrogen bonds, and desolvation (EVHDS) | -0.0053 | 0.06 | 0.18 | 0.03 |
| Energy due to electrostatic forces (EES) | 0.2833 | -0.88 | 2.57 | 0.52 |
| Energy from rotatable torsions (ETOR) | -0.0070 | -0.07 | 0.12 | 0.02 |
| Radius of gyration (Rg) | 0.2162 | 1.54 | 1.45 | 0.29 |

Table 2. Energy function parameter weights and average magnitudes of corresponding contributions to energy. Descriptor parameter weights for energy functions and mean, range, and standard deviation of the contributions of each descriptor to computed total interface energy over the training set of 78 PDBBind peptide-protein interfaces.

Figure 11 shows the distribution of peptide lengths (number of residues) in the dataset by three measures: total length of each peptide as determined from the SEQRES lines of the underlying PDB file (SRlen); the length of the segment for which coordinates are given in the PDB structure (Plen); and the length of the segment actually in contact with the I-site (Bndlen). The distribution of peptide lengths in the PPRMint dataset is highly skewed toward the shorter end, with the majority of peptides having lengths in the 8 to 14 residue range. The unbound tails (the peptide residues not in actual contact with the target) account for 19.4% of the residues of the peptides (average over 941 redundancy groups).

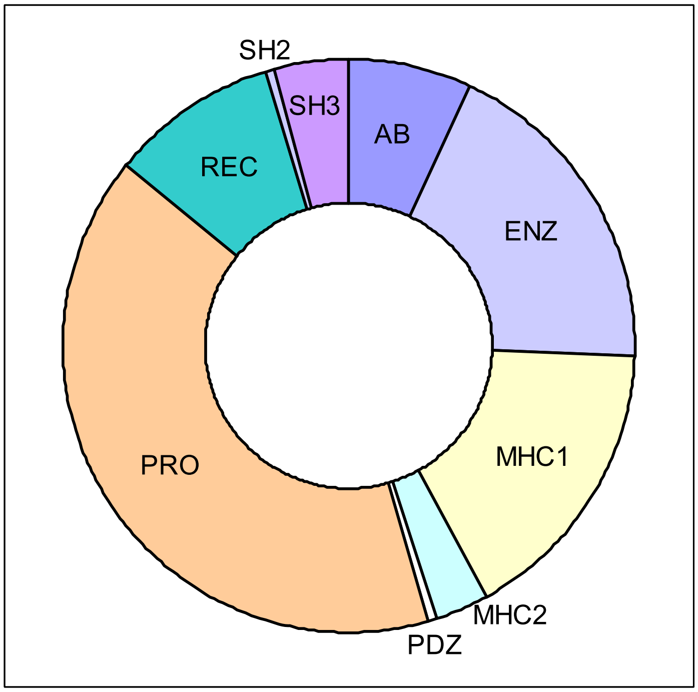

Figure 10. Interfaces by type.

The 3,924 peptides together comprise 65,850 residues. The Residues table of the PPRMint database contains records for each of the 61,130 residues for which coordinates are present in the underlying PDB structures; each of these is assigned to one of 14,592 redundancy groups (Rgrp), each containing residues that are similarly situated in terms of their surroundings, both with respect to the peptide to which they belong and the composition of the I-site region with which they are in contact  The estimated binding energies fall in a range from approximately -4 to -20 kcal/mol, with approximately 80% of the distribution falling in the -6 to -10 range (see Figure 12).

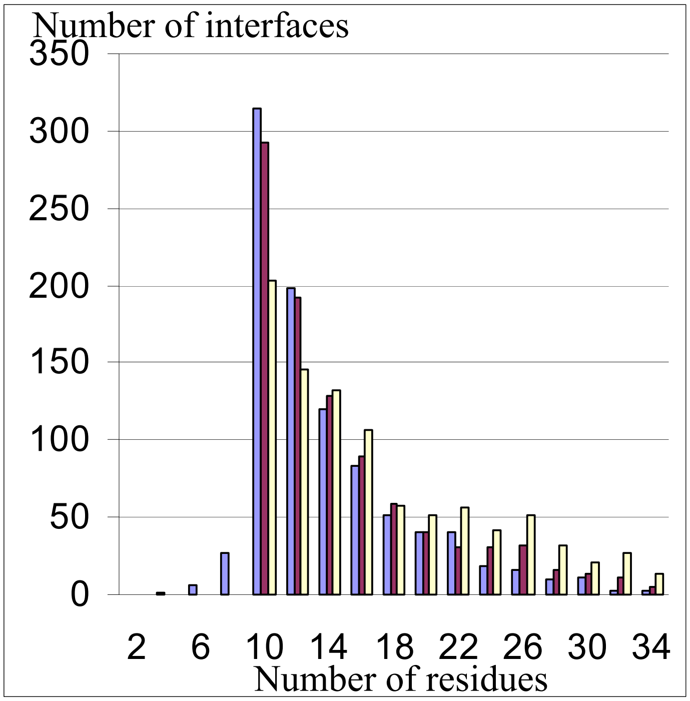

Figure 11.  Distribution of peptide lengths and contact lengths. Histogram of peptide lengths (number of residues) as reflected in PDB SEQRES lines (SRlen) (yellow), residues for which coordinates exist in PDB file (Plen) (red), and span of residues in contact with protein (Bndlen) (blue). (Data for 941 interface redundancy groups.)

The distribution of isoelectric points of the peptides (based on the full sequence as shown in the SEQRES records) is bi-modal, favoring sequences that are either strongly negatively charged or moderately positively charged, as shown in Figure 13.

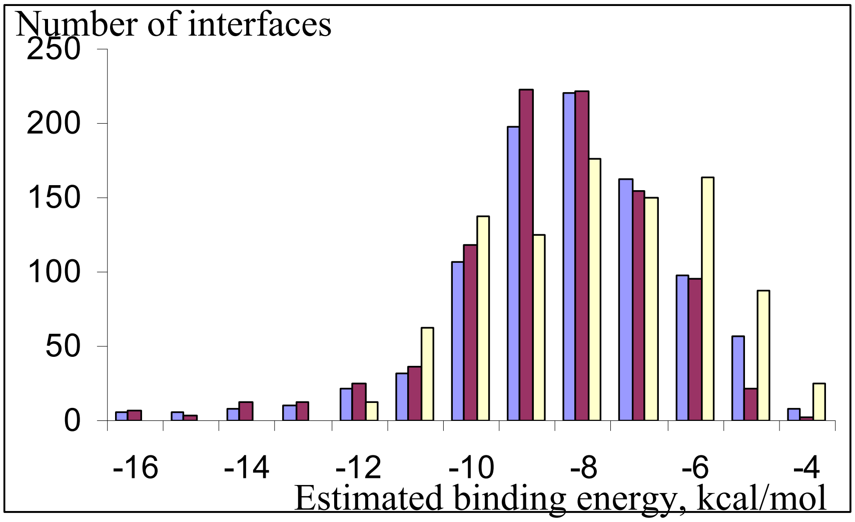

Figure 12. Distribution of estimated energies**.** Histogram of binding energies for 941 interface redundancy groups as computed by PopTop model (blue) and Autodock-based model (red); also shown are energies obtained from PDBBind database for the 75 distinct interfaces for which such data is available (yellow; counts rescaled by 941/75 to facilitate comparison).

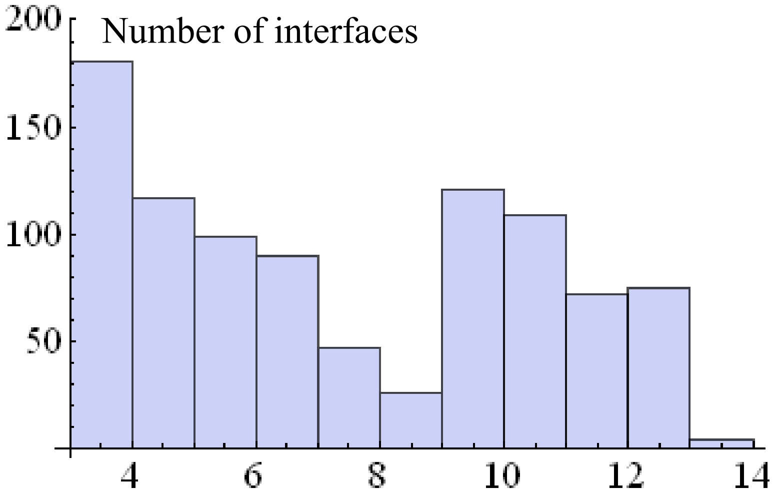

Figure 13. Peptide isoelectric points. Histogram of average isoelectric points for peptides representing 941 redundancy groups (Rgrp), based on full sequences as reported in SEQRES lines of PDB file.

The distribution of positive and negative charges within individual peptides was quite heterogeneous, with many peptides containing several positively and negatively charged residues, as shown in Figure 14. Essentially no correlation was seen between the heterogeneity / homogeneity of charge in the peptides and the estimated binding energies.  At the residue level, charged residues were, as expected, paired with oppositely charged residues or nominally uncharged residues in the great majority of cases (48% and 47% of pairings, respectively).

Surprisingly, however, as shown in Figure 15, although there was some tendency for each peptide and its I-site to be opposite in net charge, in many of the interfaces this was not the case, and, even more surprisingly, there was little correlation ($r_P = 0.13$) between the estimated binding energy and the product of the peptide net charge and I-site net charge.

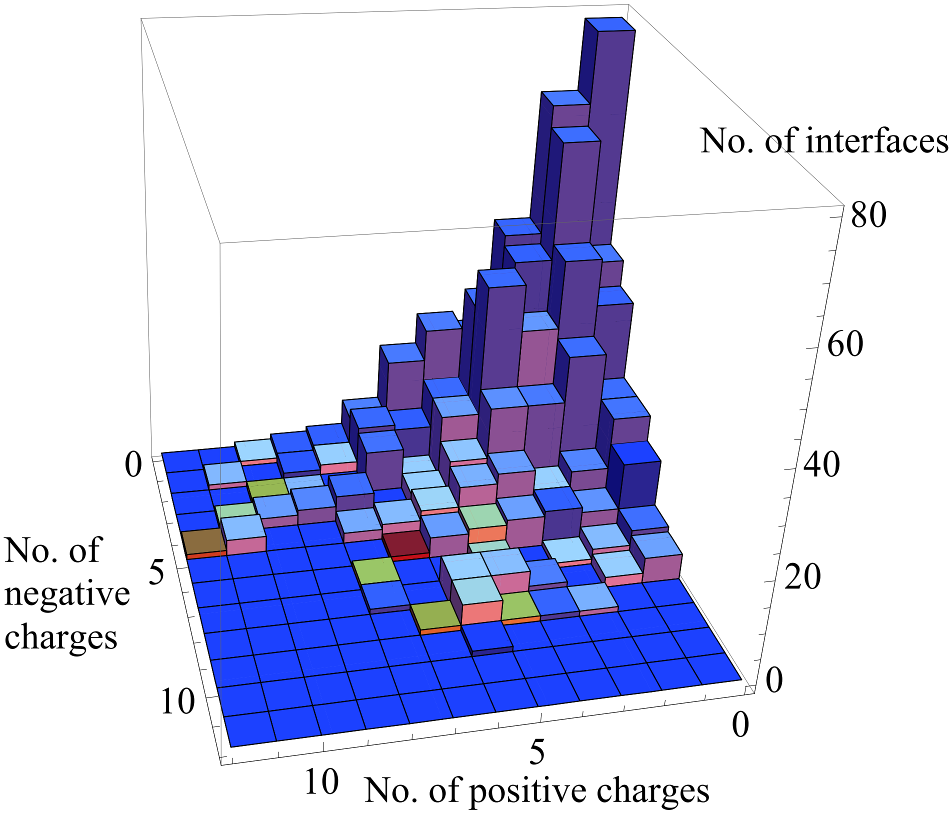

Figure 14. Distribution of charge in peptides. Number (vertical axis) of peptides having counts of positive charges (K, R, and N-terminal residue) and negative charges (D, E, and C-terminal residue) as shown (averages by group over 941 redundancy groups, residue composition as per SEQRES lines). Color indicates estimated binding energy (PopTop model), on scale ranging from 0 (blue) to -20 kcal/mol (red).

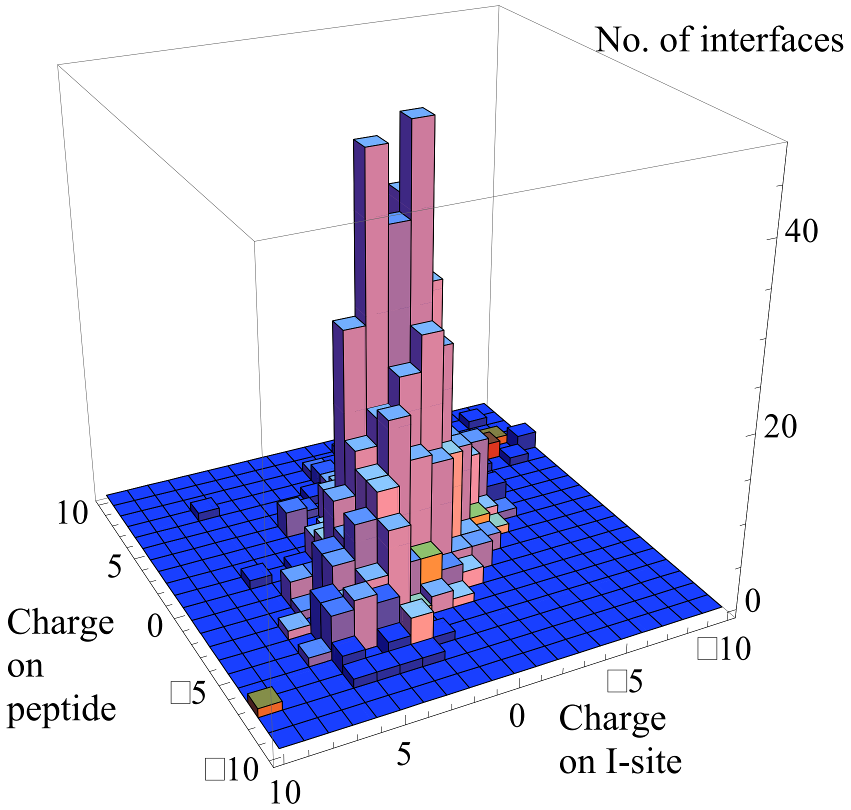

Figure 15. Distribution of charges in interfaces. Number (vertical axis) of interfaces in which the charge on the peptide and the charge on the I-site is as shown (averages by group over 941 redundancy groups). Color indicates estimated binding energy (PopTop model), on scale ranging from 0 (blue) to -20 kcal/mol (red).

In the peptides in the PPRMint dataset, shown in Figure 16 (blue bars), by comparison to residue frequencies in vertebrates generally, arginine, leucine, and proline are the most over-represented and cysteine and valine are the most underrepresented. In the residues comprising the I-sites (Figure 16, red bars), the standouts are the aromatics, with tryptophan and tyrosine fully three-fold more abundant than in vertebrates generally. As Figure 17(a) shows, by comparison to frequencies reported by others for datasets of peptide-protein interfaces and protein-protein interfaces, the peptides in the PPRMint dataset have a somewhat higher abundance of proline, arginine, lysine, and glutamate. It is common for peptides to have unbound 'tails' at either end that are not in close contact with the protein surface, as noted, and these account for about 20% of the residues, on average. The residue composition of these unbound tails differs considerably from bound portions, with serine, glycine, and the charged residues arginine, lysine, and glutamate much more abundant in the unbound tails and the hydrophobic residues isoleucine and leucine and the aromatics phenylalanine, tyrosine, and tryptophan more abundant in the bound portions.

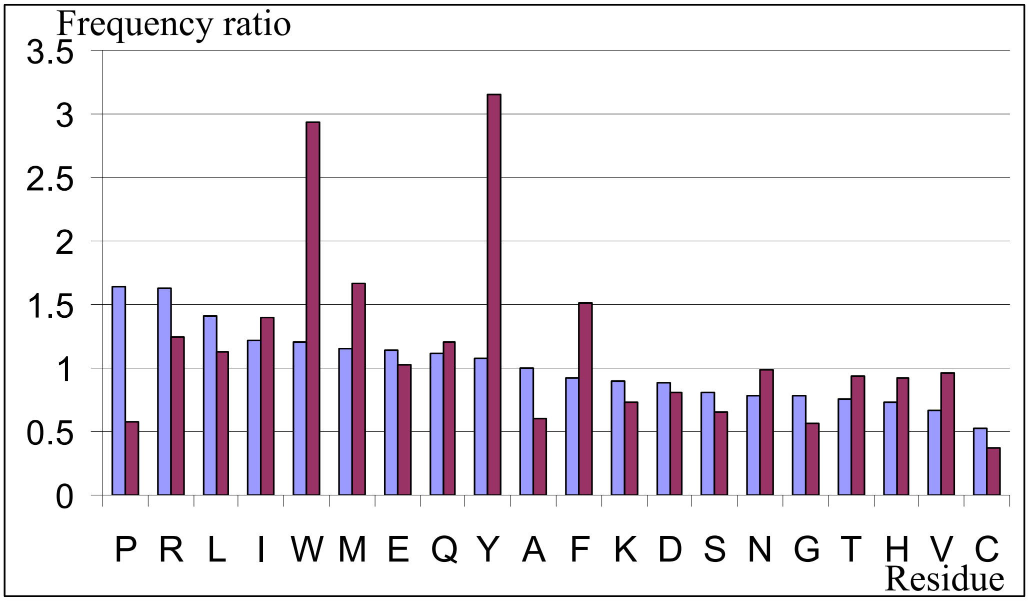

Figure 16. Amino acid frequencies in peptides and I-sites relative to frequencies in vertebrates generally. Ratio of amino acid frequencies for peptides (blue) and I-sites (red) for 941 interface redundancy groups (Rgrp), to amino acid frequencies in vertebrates generally. Peptide frequencies are based on full sequences as reported in SEQRES lines of PDB files.

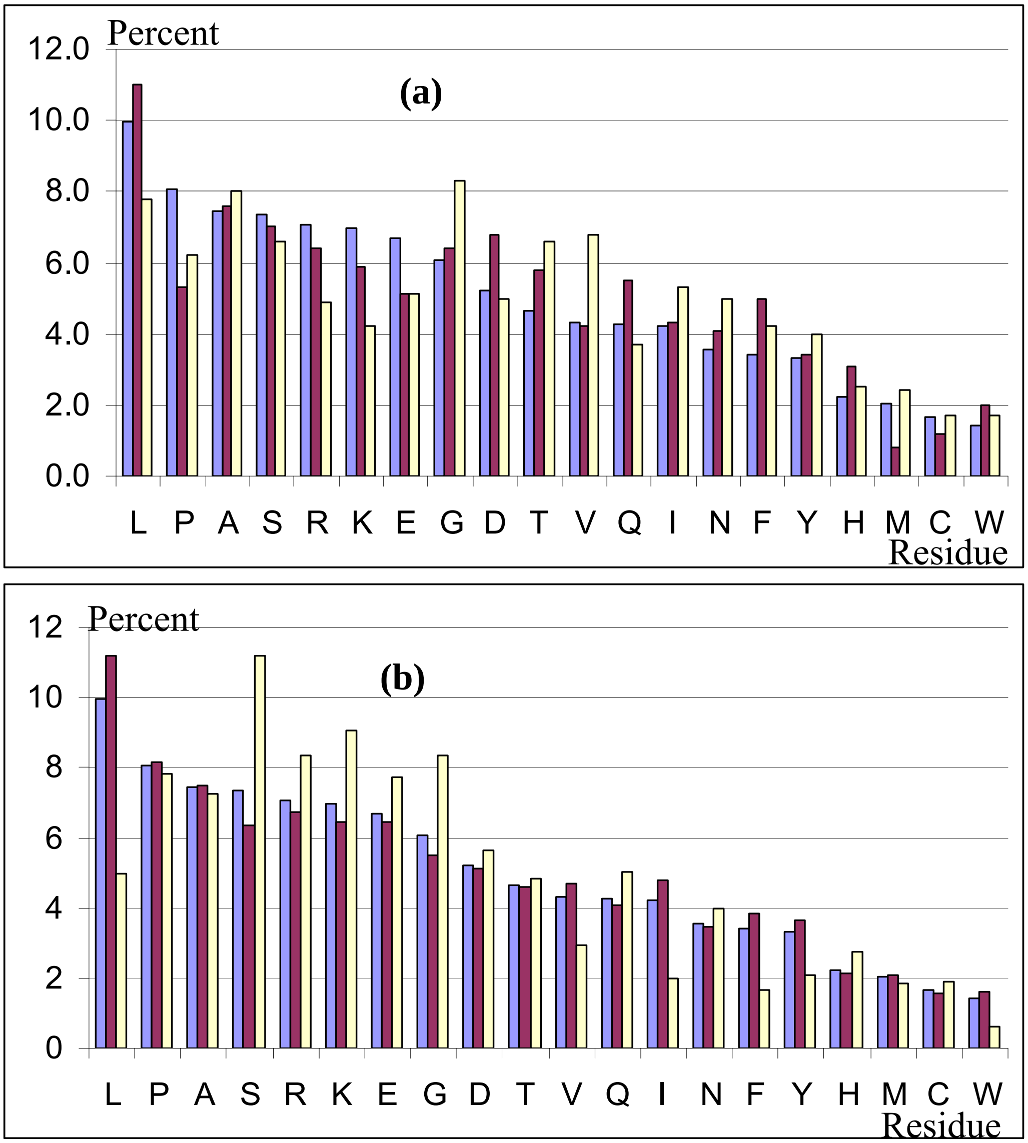

Figure 17. Amino acid frequencies in peptides. Histogram of amino acid frequencies (percent) for peptides of 941 redundancy groups (Rgrp), based on full sequences as reported in SEQRES lines of PDB files (blue), compared to: (a) frequencies reported by London et al. for peptides in another dataset of 100 peptide-protein interfaces (red) and by Glaser et al. for protein-protein interfaces (yellow); and (b) frequencies in the bound (red) and unbound (yellow) portions of the chains.

Given the complex topology of peptide-protein interfaces, it would be inaccurate to characterize binding in terms of particular peptide residues interacting with particular target residues, even if the mobility of the interface were ignored. Usually, each peptide residue has atoms within the Bjerrum length of atoms from multiple target residues. Nevertheless, it is instructive to ask whether there is a higher than random tendency for certain residue types to be present in the I-site at or near particular peptide residue types. Table 3 shows, for the 14,592 non-redundant clusters of peptide residues, the number of I-site residues of each amino acid type that were nearest to each peptide amino acid type (as measured by the distance between the closest atoms of each). By this measure, the most frequent pairings were of the positively charged residues R and K in the peptide with the negatively charged residues D and E in the I-site. Other frequent pairings were P in the peptide with W and Y in the I-site, and L in the peptide with K or L in the I-site.

The bound conformations of the peptides in the PPRMint dataset are nearly all relatively extended, with the actual end-to-end length of the peptides exceeding the theoretical rms end-to-end length for a random flight conformation (equal to the distance between alpha carbons, 3.8Å, times $\sqrt{n-1}$, where n is the number of residues [51, 61]) by an average ratio of 1.76; the peptides in 845 of the 941 interface groups are at or above theoretical rms random flight conformation length (Figure 18).

The secondary structure tendencies of the bound conformations of the peptides were evaluated by categorizing each residue into one of three mesostates – helical (‘a’), including α, $3_{10}$, and π helices; extended β strand (‘b’); or left-handed helical (‘l’) – based on backbone Ramachandran angles, according to the method of Ho and Dill [68]. As shown in Figure 19(a), most peptides had 10% or fewer residues in the left-handed helical compartment of Ramachandran space, and residues categorized as helical (including α, $3_{10}$, or π forms) and extended β-strand comprised, respectively, percentages of total peptide composition that were distributed fairly evenly over the range from 0 to 100%, with β-strand predominating. Overall, 44.6% of the residues of the peptides in the 941 interface groups were categorized as helical, 51.1% as extended β-strand, and 4.3% as left-handed helical.

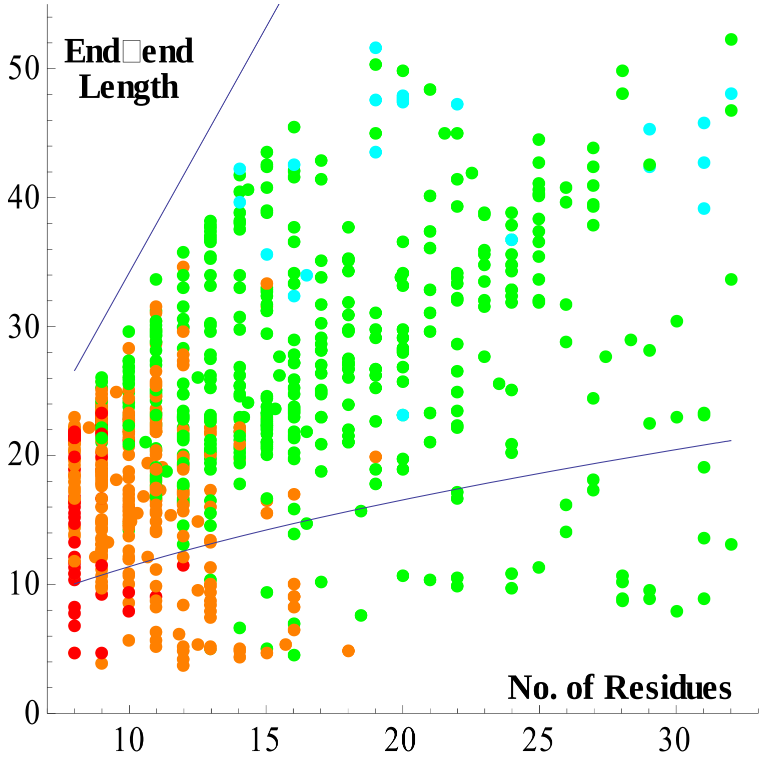

Figure 18. Peptide actual lengths vs. random flight lengths. End-to-end lengths of peptides in interfaces (distance between terminal Cα atoms, Ǻ) vs. number of residues, averages for each of 941 redundancy groups. Colors indicate estimated binding energy per residue according to PopTop model: red, energy >= 75% of maximum in set; orange, >= 50%; green, >= 25%; cyan, lowest 25%.  Lower line: theoretical random flight length for number of residues present. Upper line: maximum possible fully extended length.

As Figure 19(b) shows, extended β-strand residues often occur in contiguous stretches spanning 50% or more of the length of the peptide; helical segments are typically somewhat shorter.  This data again suggests that many of the peptides are bound in relatively extended conformations generally lacking in internal structure.

| Pres | All | Unb | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 1071 | 180 | 22 | 7 | 27 | 70 | 40 | 13 | 25 | 48 | 73 | 72 | 27 | 50 | 23 | 48 | 80 | 28 | 32 | 78 | 50 | 79 |
| **C** | 248 | 72 | 5 | 28 | 6 | 7 | 4 | 14 | 3 | 4 | 5 | 9 | 3 | 10 | 2 | 16 | 18 | 5 | 8 | 11 | 7 | 10 |
| **D** | 762 | 172 | 9 | 0 | 2 | 4 | 6 | 17 | 30 | 6 | 171 | 20 | 6 | 22 | 6 | 36 | 138 | 26 | 23 | 13 | 10 | 44 |
| **E** | 957 | 229 | 13 | 0 | 6 | 17 | 17 | 11 | 19 | 15 | 154 | 20 | 18 | 29 | 13 | 48 | 162 | 39 | 21 | 29 | 19 | 76 |
| **F** | 501 | 31 | 17 | 1 | 7 | 19 | 22 | 14 | 18 | 18 | 37 | 32 | 9 | 31 | 20 | 28 | 30 | 26 | 28 | 30 | 24 | 58 |
| **G** | 817 | 169 | 26 | 2 | 27 | 45 | 41 | 17 | 24 | 21 | 78 | 32 | 17 | 36 | 19 | 46 | 59 | 39 | 23 | 16 | 32 | 51 |
| **H** | 342 | 39 | 11 | 5 | 20 | 50 | 5 | 5 | 2 | 19 | 20 | 27 | 7 | 14 | 8 | 13 | 12 | 27 | 10 | 19 | 10 | 18 |
| **I** | 658 | 53 | 18 | 9 | 7 | 70 | 30 | 23 | 17 | 46 | 33 | 60 | 24 | 25 | 2 | 26 | 30 | 24 | 39 | 36 | 15 | 72 |
| **K** | 992 | 183 | 24 | 1 | 136 | 197 | 23 | 22 | 37 | 27 | 37 | 28 | 13 | 42 | 19 | 39 | 29 | 29 | 20 | 21 | 27 | 40 |
| **L** | 1488 | 119 | 45 | 5 | 31 | 122 | 56 | 34 | 23 | 75 | 221 | 145 | 38 | 64 | 33 | 63 | 87 | 42 | 59 | 79 | 42 | 104 |
| **M** | 342 | 37 | 14 | 0 | 6 | 30 | 20 | 7 | 12 | 21 | 29 | 29 | 5 | 18 | 8 | 19 | 19 | 11 | 13 | 10 | 13 | 20 |
| **N** | 495 | 82 | 7 | 5 | 23 | 41 | 28 | 15 | 8 | 12 | 40 | 14 | 11 | 36 | 8 | 34 | 38 | 15 | 17 | 14 | 15 | 30 |
| **P** | 1166 | 154 | 13 | 11 | 33 | 49 | 64 | 30 | 20 | 36 | 50 | 31 | 14 | 66 | 39 | 47 | 62 | 48 | 48 | 37 | 136 | 178 |
| **Q** | 637 | 126 | 25 | 3 | 29 | 37 | 19 | 11 | 24 | 15 | 50 | 41 | 16 | 29 | 11 | 39 | 47 | 24 | 26 | 23 | 13 | 31 |
| **R** | 1111 | 184 | 20 | 5 | 223 | 276 | 25 | 14 | 18 | 17 | 30 | 26 | 16 | 36 | 5 | 43 | 31 | 26 | 29 | 27 | 21 | 38 |
| **S** | 970 | 201 | 36 | 6 | 64 | 112 | 15 | 19 | 26 | 15 | 95 | 33 | 16 | 30 | 16 | 59 | 56 | 24 | 35 | 39 | 25 | 49 |
| **T** | 679 | 116 | 14 | 5 | 45 | 37 | 26 | 16 | 12 | 24 | 46 | 20 | 6 | 33 | 20 | 35 | 58 | 23 | 16 | 38 | 42 | 47 |
| **V** | 637 | 54 | 24 | 3 | 11 | 25 | 23 | 20 | 24 | 46 | 39 | 63 | 16 | 35 | 16 | 33 | 28 | 24 | 41 | 38 | 24 | 49 |
| **W** | 198 | 17 | 5 | 2 | 2 | 12 | 9 | 10 | 7 | 9 | 10 | 22 | 16 | 8 | 5 | 11 | 14 | 11 | 9 | 8 | 4 | 7 |
| **Y** | 520 | 48 | 3 | 2 | 32 | 29 | 11 | 12 | 30 | 26 | 42 | 23 | 18 | 28 | 24 | 28 | 43 | 27 | 13 | 24 | 23 | 33 |
| **All** | | 2266 | 351 | 100 | 737 | 1249 | 484 | 324 | 379 | 500 | 1260 | 747 | 296 | 642 | 297 | 711 | 1041 | 518 | 510 | 590 | 552 | 1034 |

Table 3. Relative prevalence of non-redundant pairings of peptide residues with target residues. Counts of target residue type in closest proximity to peptide residue for 14,591 non-redundant pairings. Peptide residues are in left-most column, columns identified in top row indicate the target protein residue type. “Unb” column contains the counts for peptide residues that have no target residue with any atoms within 4Å of any atoms of the peptide residue. Mean count: 30.8; standard deviation (σ):32.9. Blue: more than 0.5σ below mean; grey: between -0.5σ and + 0.5σ; yellow: between +0.5σ and 3σ; orange: above 3σ.

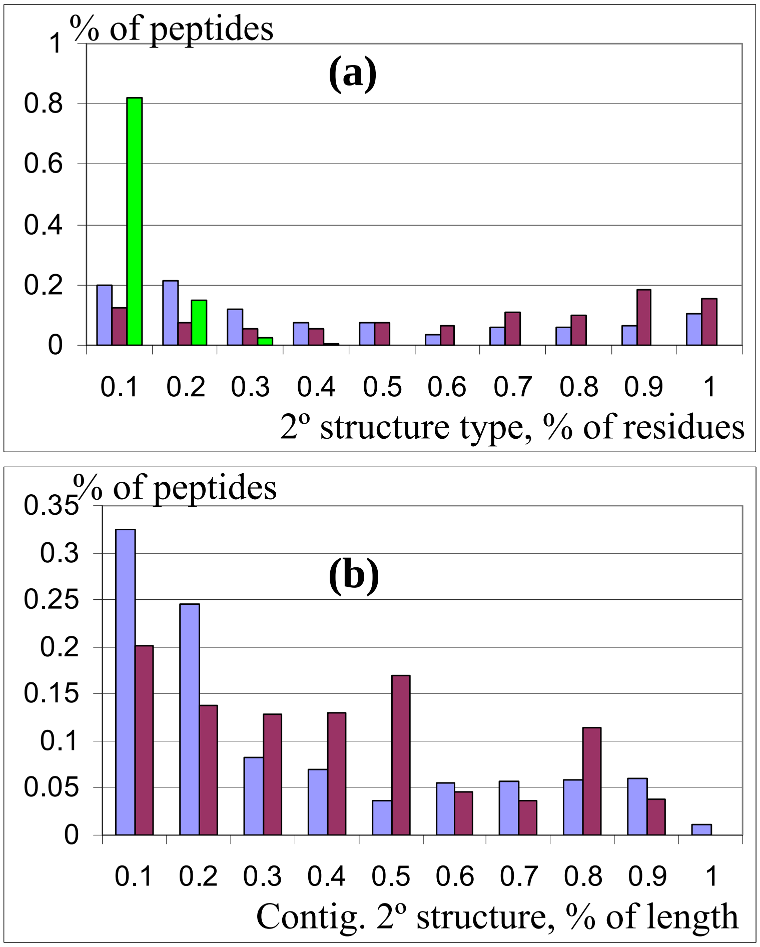

Figure 19. Distributions of secondary structure tendencies in peptides. Percentages (vertical axis) of peptides in interfaces having (a) percentages shown (horizontal axis) of helical (blue), extended β-strand (red), and left-handed helical (green) residues, and (b) wherein the longest contiguous subsequence of helical (blue) and β-strand (red) residues comprised the percentage shown (horizontal axis) of the total length of the peptide (averages by group over 941 redundancy groups).

### Variation in bound position within redundancy groups

To assess the extent to which the position of the bound peptide varies relative to the protein surface in essentially identical interfaces, a pairwise comparison was made of binding 'signatures' (as described above) for each pair of interfaces in each of the 150 interface redundancy groups (Rgrp) that contain five or more interfaces. With the peptide residues in each pair of signatures aligned for optimal matchup, for each pair of peptide residues in the pair of aligned signatures, a count was made of the number of proximal (within 4Ǻ) I-site residues associated with one signature and not the other. The resulting count, taken over all the peptide residues in the pair of signatures, and divided by the total number of proximal protein residues associated with either peptide, gives a measure of the fraction of I-site residues that are different in the signatures of the two interfaces. By populating a square difference matrix with these difference fractions, selecting the column having the minimum L1-norm as representing the most compact representation of the cluster, and taking the mean of the pairwise differences comprising the elements of the column, a measure was obtained that may be thought of as a kind of 'radius' of the cluster, and representative of the average percent variability in position of the peptides in the redundancy group. As shown in Figure 20(a), in approximately half of these redundancy groups, there is approximately ten percent or more average difference between the sets of I-site residues in proximity to the peptide in any pair of interfaces in the group. The variability of estimated binding energy was also evaluated among interfaces belonging to the same redundancy group; as Figure 20(b) shows, typical coefficients of variation of estimated binding energy are in the range of 5 to 10 percent.

### Relationship of gross geometric characteristics to estimated binding energy

On average, the binding energies as estimated by both energy models increase monotonically with length over the range of peptide lengths represented in the dataset, with the average per-residue energy contribution levelling off at approximately -0.4 kcal/mol for chain lengths of about 20 or above, as shown in Figure 21.

Interfaces in which the peptide occupies a concavity in the target tend to have correspondingly larger estimated binding energies than those in which the peptide is more superficially associated, as shown in Figure 22(a), which plots the estimated binding energy against the ratio of the peptide's ASA in the bound interface to its ASA in the same conformation but without the presence of the target ($r_P$ = 67%). Similarly, there is a strong correlation between estimated binding energy and the ratio of the number of residues in the I-site to the number of residues in the peptide, implying that when the peptide is 'surrounded' by a larger number of I-site residues, improved energy may be expected (Figure 22(b)).

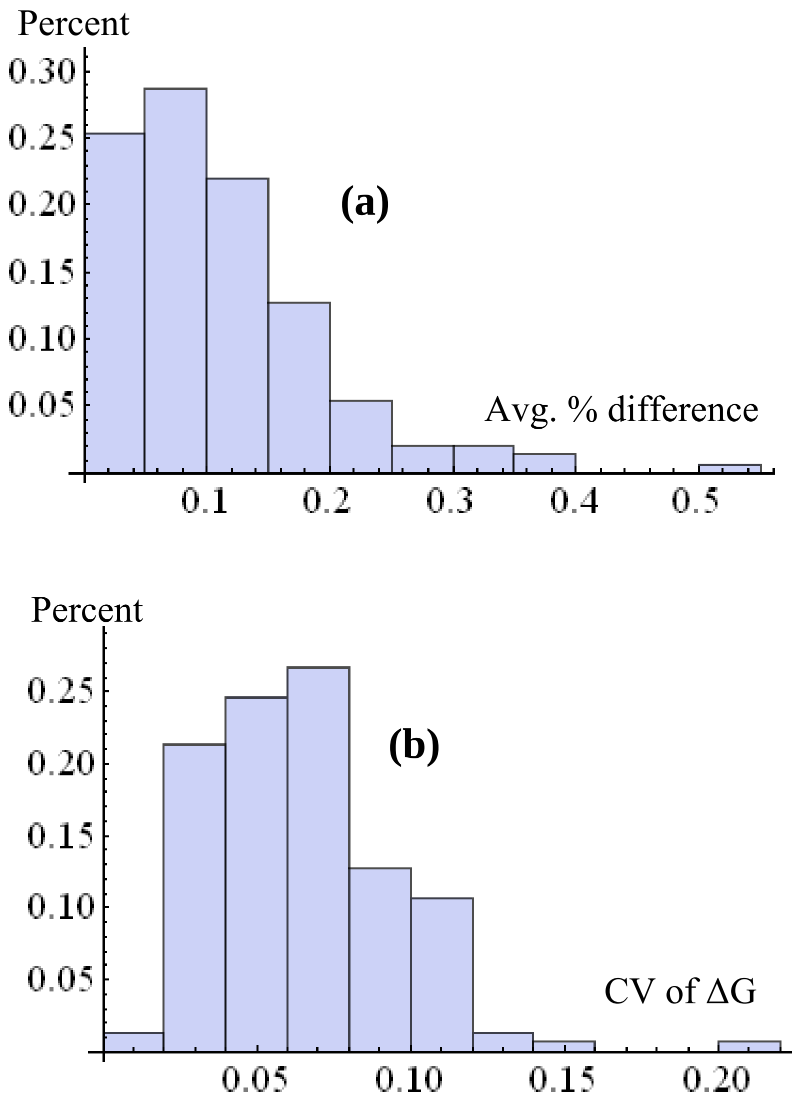

Figure 20. Variability in peptide position relative to protein surface residues. Frequency histograms of (a) redundancy groups by average fraction of residues within 4A of each peptide residue differing between any two interfaces in the redundancy group; (b) redundancy groups by coefficients of variation of predicted binding energy (PopTop model), for 150 redundancy groups (Rgrp) containing at least five members.

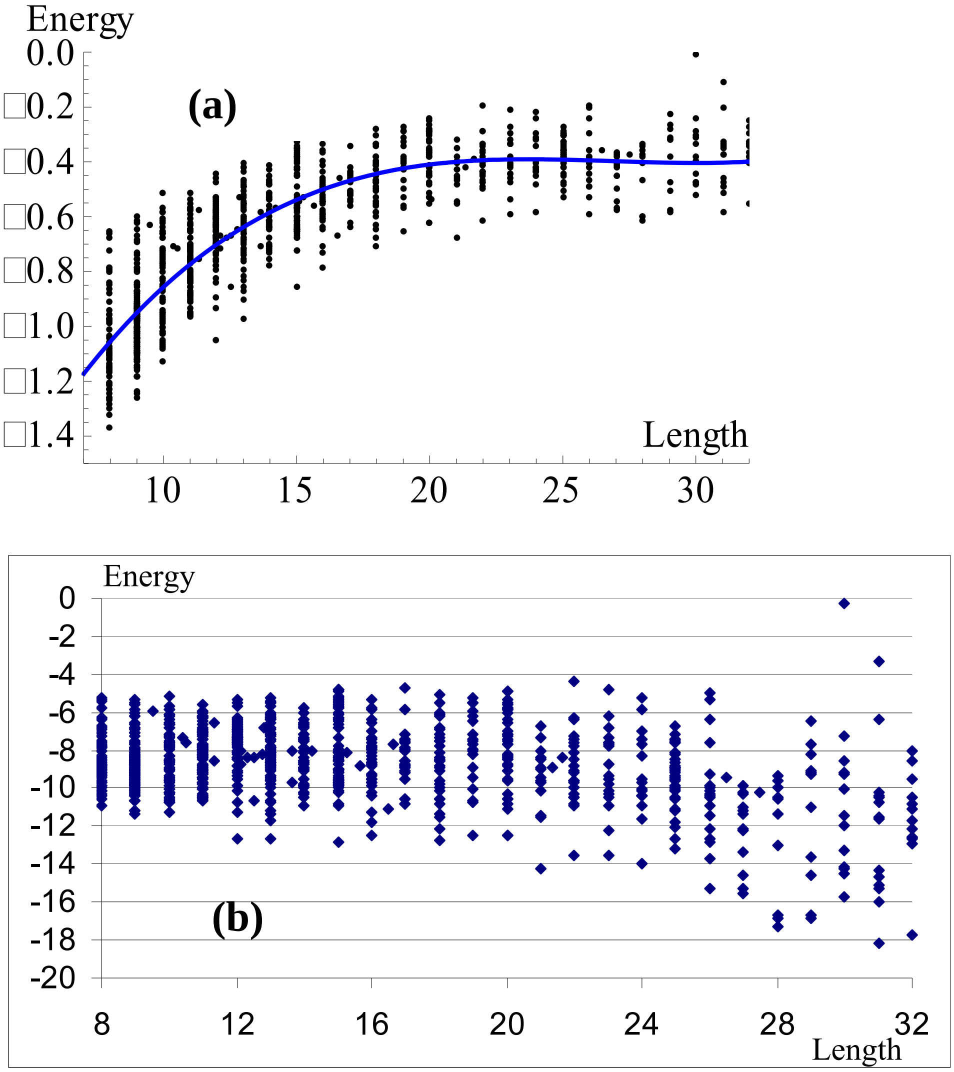

Figure 21. Binding energy vs. peptide length. Binding energies (a) per unit peptide length (line is least squares fit to cubic) and (b) for entire interface, vs peptide length (as determined from SEQRES). Energies as estimated by PopTop model, averages over 941 redundancy groups (Rgrp).

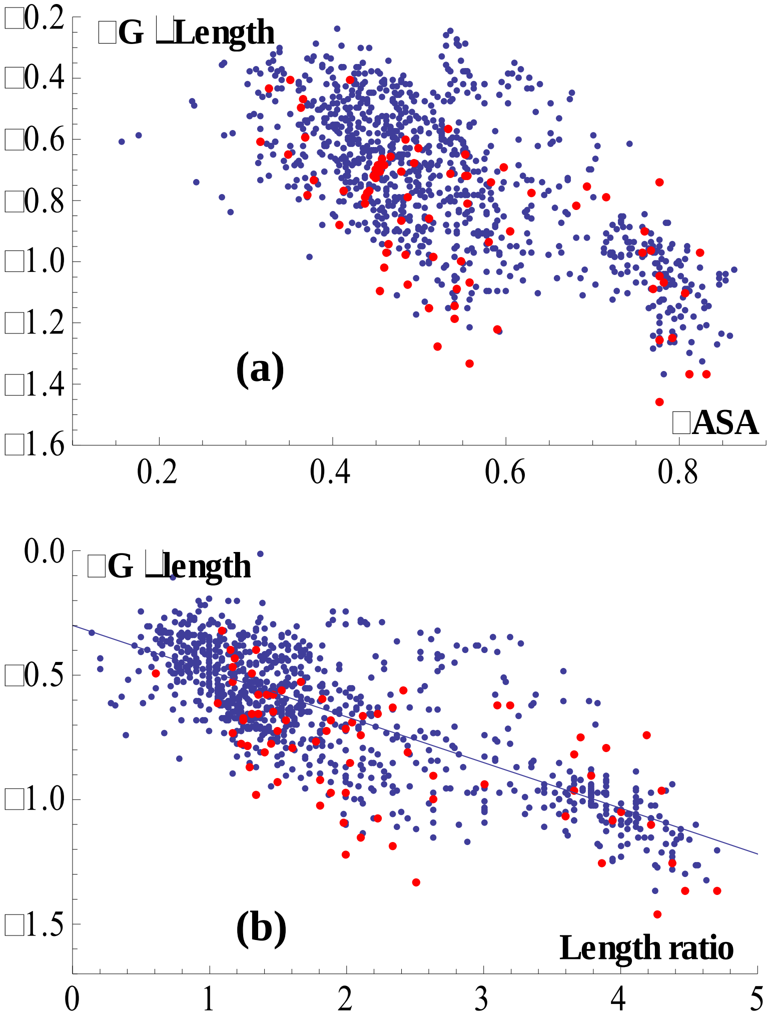

Figure 22. Estimated binding energy vs embeddedness. Binding energy per residue vs. (a) percentage of peptide's accessible surface area that is no longer accessible in the complex with protein; (b) ratio of number of residues in I-site to number of residues in peptide. Averages for each of 941 redundancy groups. Blue: energies as estimated by PopTop model. Red: energies from PDBBind database for 75 redundancy groups of interfaces present in PDBBind database.

For the interfaces of the PPRMint dataset, no significant gross relationship was seen between chain extendedness and estimated binding energy ($r_P$ = 0.15); however, as Figure 23 makes apparent, the peptides having the highest estimated per-residue binding energy form a cluster characterized by actual end to end lengths of approximately double the theoretical rms length for random flight conformations.

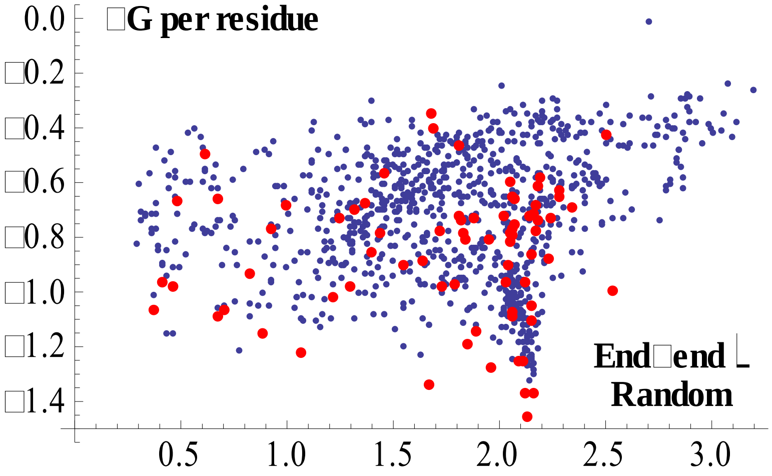

Figure 23. Estimated binding energy vs ratio of actual length to random flight length. Binding energy per residue vs. ratio of end-to-end lengths of peptides in interfaces (distance between terminal Cα atoms, Å) to theoretical random flight length (Å), averages for each of 941 redundancy groups. Blue: energies as estimated by PopTop model. Red: energies from PDBBind database for 75 redundancy groups of interfaces present in PDBBind database.

To assess the relationship, if any, between peptide secondary structure tendencies and binding energy, the interface binding energies as estimated by the PopTop model (averaged over each interface redundancy group (Rgrp)) were compared with the percentages of each peptide sequence categorized (by mesostate, as described above) as having helical, extended β-strand, and left-handed helical tendencies, and with the lengths of the longest contiguous helical and β-strand subsequences in each peptide. Little correlation was seen between the helical, β-strand, and left-handed helix percentages and the estimated binding energy per residue ($r_P$ = 0.26, -0.27, and 0.06, respectively), even after normalizing the estimated binding energies to eliminate the dependence on peptide length by dividing the estimated energy per residue by the value given by the regression line in Figure 21(a) ($r_P$ = 0.32, -0.33, and 0.06, respectively). There did appear to be some negative relationship between the length of the longest beta strand as a percentage of peptide length, and the estimated binding energy per residue normalized for length effects ($r_P$ = 0.44); as is evident from Figure 24, the likelihood of having a higher than average per-residue binding energy declines noticeably if the longest beta strand segment occupies more than about half of the sequence.

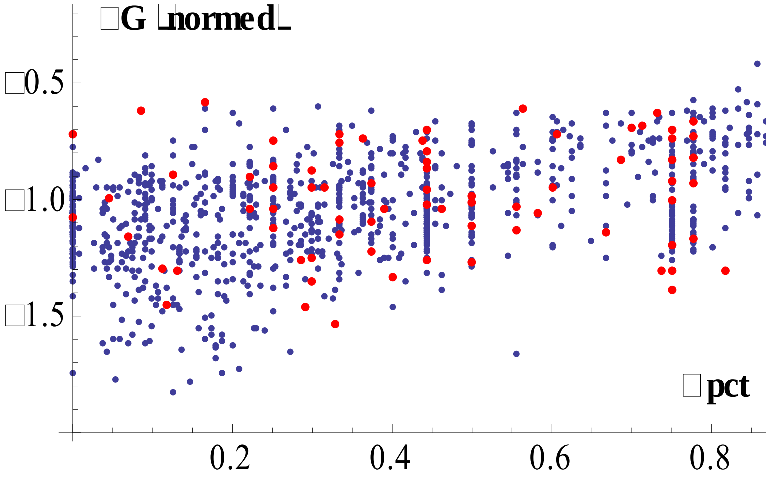

Figure 24. Estimated binding energy per residue vs maximum beta strand length. Binding energy per residue, normalized to eliminate dependence on peptide length, vs. length of longest contiguous beta strand as percentage of peptide length, averages for each of 941 redundancy groups. Blue: energies as estimated by PopTop model. Red: energies from PDBBind database for 75 redundancy groups of interfaces present in PDBBind database.

### Relationship of peptide composition to binding energy

The contributions of each peptide residue to binding energy were estimated based on both the PopTop model (Table 4) and the Autodock-based model (Table 5), with the energy contributions further allocated according to the amino acid type of the nearest neighboring residue in the I-site. (Data reflects averages for each category over the 14,592 non-redundant residue pairings as clustered in the (Rgrp) residue redundancy groups.) Results from both models identify generally the same peptide residues and residue pairings as making the greatest and least energy contributions; both attribute relatively high energies to charge-charge interactions and aromatics.

| Pres | Avg |  | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | -0.73 | 0.18 | -0.65 | -0.79 | -0.85 | -0.86 | -0.80 | -0.87 | -0.75 | -0.69 | -0.77 | -0.72 | -0.80 | -0.74 | -0.76 | -0.74 | -0.72 | -0.75 | -0.79 | -0.71 | -0.80 | -1.03 |
| C | -0.82 | 0.10 | -0.57 | -0.75 | -0.90 | -0.96 | -0.84 | -0.97 | -0.80 | -0.81 | -0.83 | -1.04 | -0.78 | -0.70 | -1.23 | -0.96 | -0.82 | -0.88 | -0.81 | -0.92 | -0.70 | -0.96 |
| D | -0.68 | 0.21 | -0.81 |  | -0.69 | -0.57 | -0.69 | -0.76 | -0.76 | -0.60 | -0.92 | -0.74 | -0.74 | -0.71 | -0.70 | -0.72 | -1.00 | -0.82 | -0.75 | -0.77 | -0.73 | -1.01 |
| E | -0.66 | 0.25 | -0.75 |  | -0.67 | -0.76 | -0.54 | -0.77 | -0.73 | -0.59 | -0.91 | -0.81 | -0.56 | -0.82 | -0.57 | -0.71 | -0.97 | -0.90 | -0.80 | -0.63 | -0.71 | -1.00 |
| F | -0.83 | 0.23 | -0.81 | -1.30 | -0.62 | -0.80 | -0.85 | -0.82 | -0.85 | -0.79 | -0.98 | -0.96 | -0.74 | -0.89 | -0.89 | -0.79 | -0.86 | -0.92 | -1.06 | -0.84 | -0.76 | -1.09 |
| G | -0.76 | 0.17 | -0.74 | -0.66 | -0.86 | -0.89 | -0.84 | -0.94 | -0.80 | -0.80 | -0.81 | -0.74 | -0.82 | -0.77 | -0.67 | -0.82 | -0.72 | -0.82 | -0.83 | -0.74 | -0.85 | -1.02 |
| H | -0.71 | 0.25 | -0.67 | -0.92 | -0.82 | -1.07 | -0.62 | -0.22 | -0.86 | -0.70 | -0.88 | -0.89 | -0.65 | -0.55 | -0.95 | -0.71 | -0.65 | -0.93 | -0.64 | -0.97 | -0.57 | -0.82 |
| I | -0.75 | 0.23 | -0.74 | -1.00 | -0.82 | -0.94 | -0.86 | -0.69 | -0.80 | -0.69 | -0.89 | -0.73 | -0.83 | -0.56 | -1.03 | -0.80 | -0.79 | -0.82 | -0.74 | -0.76 | -0.76 | -0.80 |
| K | -0.67 | 0.26 | -0.62 | -0.65 | -1.18 | -1.08 | -0.78 | -0.77 | -0.46 | -0.50 | -0.70 | -0.60 | -0.74 | -0.73 | -0.51 | -0.58 | -0.63 | -0.69 | -0.65 | -0.60 | -0.57 | -1.26 |
| L | -0.76 | 0.19 | -0.76 | -0.81 | -0.82 | -0.91 | -0.81 | -0.76 | -0.73 | -0.80 | -0.89 | -0.78 | -0.79 | -0.77 | -0.79 | -0.81 | -0.76 | -0.80 | -0.95 | -0.78 | -0.67 | -1.03 |
| M | -0.63 | 0.27 | -0.75 |  | -0.78 | -0.90 | -0.60 | -0.68 | -0.62 | -0.74 | -0.77 | -0.51 | -0.58 | -0.77 | -0.61 | -0.75 | -0.75 | -0.71 | -0.71 | -0.69 | -0.69 | -0.89 |
| N | -0.70 | 0.24 | -0.44 | -0.94 | -0.85 | -0.87 | -0.70 | -0.69 | -0.88 | -0.72 | -0.75 | -0.60 | -0.80 | -0.77 | -0.74 | -0.73 | -0.62 | -0.78 | -0.70 | -0.68 | -0.79 | -0.78 |
| P | -0.73 | 0.24 | -0.78 | -0.84 | -0.87 | -0.87 | -0.75 | -0.76 | -0.74 | -0.75 | -0.69 | -0.89 | -0.85 | -0.79 | -0.76 | -0.74 | -0.74 | -0.79 | -0.77 | -0.64 | -0.74 | -0.76 |
| Q | -0.64 | 0.26 | -0.69 | -0.71 | -0.71 | -0.81 | -0.73 | -0.47 | -0.68 | -0.52 | -0.64 | -0.62 | -0.80 | -0.83 | -0.59 | -0.59 | -0.68 | -0.60 | -0.88 | -0.53 | -0.69 | -0.84 |
| R | -0.69 | 0.23 | -0.78 | -0.77 | -1.27 | -1.16 | -0.82 | -0.71 | -0.57 | -0.63 | -0.47 | -0.63 | -0.67 | -0.71 | -0.52 | -0.65 | -0.66 | -0.63 | -0.67 | -0.61 | -0.74 | -0.99 |
| S | -0.75 | 0.21 | -0.56 | -0.86 | -0.84 | -0.90 | -0.86 | -0.74 | -0.76 | -0.66 | -0.80 | -0.88 | -0.77 | -0.79 | -0.91 | -0.64 | -0.84 | -0.76 | -0.70 | -0.79 | -0.81 | -1.04 |
| T | -0.70 | 0.21 | -0.62 | -0.69 | -0.81 | -0.92 | -0.70 | -0.72 | -0.64 | -0.64 | -0.76 | -0.74 | -0.71 | -0.69 | -0.75 | -0.80 | -0.67 | -0.79 | -0.79 | -0.77 | -0.75 | -0.94 |
| V | -0.70 | 0.18 | -0.60 | -0.63 | -0.85 | -0.74 | -0.64 | -0.74 | -0.68 | -0.68 | -1.01 | -0.75 | -0.60 | -0.67 | -0.74 | -0.67 | -0.79 | -0.75 | -0.84 | -0.64 | -0.77 | -1.10 |
| W | -0.84 | 0.15 | -0.73 | -0.67 | -1.24 | -0.95 | -0.82 | -0.73 | -1.01 | -0.73 | -0.80 | -0.87 | -1.08 | -0.71 | -1.31 | -0.65 | -0.99 | -0.84 | -0.87 | -0.93 | -0.76 | -1.02 |
| Y | -0.84 | 0.29 | -0.83 | -0.91 | -1.10 | -1.24 | -0.59 | -0.90 | -0.77 | -0.66 | -0.81 | -0.95 | -0.78 | -0.95 | -1.13 | -0.71 | -0.95 | -1.15 | -0.91 | -1.05 | -0.69 | -0.93 |
| Avg |  |  | -0.70 | -0.82 | -0.88 | -0.91 | -0.74 | -0.74 | -0.74 | -0.69 | -0.80 | -0.77 | -0.75 | -0.75 | -0.81 | -0.73 | -0.78 | -0.81 | -0.79 | -0.75 | -0.73 | -0.97 |

Table 4. Predicted interaction energies for non-redundant pairings of peptide residues with target residues (PopTop model). Residue-level contributions to binding energies (kcal/mol, as estimated by PopTop model) for peptide residues of the type shown in the left-most column when the target residue type in closest proximity is that shown in the top row, averages based on 14,591 non-redundant pairings.. "Unb" column contains the energies for peptide residues that have no target residue with any atoms within 4A of any atoms of the peptide residue. Mean: -0.78. Standard deviation (σ): 0.15. Blue: more than 2 σ above mean; light blue: between +1σ and +2σ; grey: between -1σ and +1σ; yellow: between -1σ and -2σ; orange: more than 2σ below mean (note energies are negative, lower value connotes higher energy interaction).

| Pres | Avg | | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | -0.85 | 0.28 | -0.80 | -0.89 | -0.98 | -0.92 | -0.93 | -0.96 | -0.85 | -0.86 | -0.94 | -0.87 | -0.97 | -0.91 | -0.89 | -0.85 | -0.81 | -0.90 | -0.93 | -0.87 | -0.95 | -1.05 |
| C | -0.94 | 0.15 | -0.54 | -1.00 | -0.99 | -1.09 | -0.94 | -1.21 | -0.93 | -0.91 | -0.98 | -1.28 | -0.88 | -0.86 | -1.49 | -1.04 | -0.85 | -0.96 | -0.92 | -1.16 | -0.82 | -1.02 |
| D | -0.70 | 0.29 | -0.77 | | -0.58 | -0.41 | -0.67 | -0.94 | -0.85 | -0.53 | -1.03 | -0.75 | -0.83 | -0.82 | -0.73 | -0.79 | -1.05 | -0.98 | -0.81 | -0.66 | -0.84 | -1.06 |
| E | -0.64 | 0.36 | -0.77 | | -0.47 | -0.58 | -0.51 | -0.74 | -0.74 | -0.52 | -0.97 | -0.78 | -0.63 | -0.86 | -0.57 | -0.72 | -0.93 | -0.98 | -0.84 | -0.59 | -0.67 | -1.03 |
| F | -0.81 | 0.36 | -0.95 | -1.17 | -0.63 | -0.74 | -0.93 | -0.72 | -0.94 | -0.93 | -0.98 | -0.90 | -0.86 | -0.89 | -0.84 | -0.71 | -0.73 | -0.86 | -0.91 | -0.93 | -0.82 | -0.99 |
| G | -0.87 | 0.27 | -0.93 | -0.71 | -0.86 | -0.97 | -0.99 | -1.08 | -0.92 | -0.95 | -0.94 | -0.84 | -0.97 | -0.94 | -0.80 | -0.95 | -0.83 | -0.95 | -1.02 | -0.91 | -0.96 | -1.07 |
| H | -0.67 | 0.38 | -0.65 | -0.93 | -0.76 | -0.94 | -0.68 | -0.24 | -0.71 | -0.66 | -0.84 | -0.92 | -0.59 | -0.58 | -0.91 | -0.63 | -0.66 | -0.93 | -0.55 | -1.01 | -0.49 | -0.74 |
| I | -0.86 | 0.35 | -0.93 | -1.17 | -0.73 | -0.91 | -0.98 | -0.91 | -0.94 | -0.92 | -1.04 | -0.95 | -1.05 | -0.69 | -1.08 | -0.87 | -0.90 | -0.88 | -0.88 | -0.95 | -0.84 | -0.84 |
| K | -0.66 | 0.39 | -0.66 | -0.47 | -1.16 | -1.05 | -0.82 | -0.90 | -0.47 | -0.54 | -0.65 | -0.69 | -0.81 | -0.76 | -0.42 | -0.57 | -0.49 | -0.73 | -0.70 | -0.72 | -0.58 | -0.97 |
| L | -0.86 | 0.28 | -0.90 | -0.84 | -0.78 | -1.05 | -0.96 | -0.76 | -0.78 | -0.95 | -1.10 | -0.97 | -1.00 | -0.82 | -0.91 | -0.90 | -0.87 | -0.84 | -1.08 | -0.97 | -0.79 | -1.12 |
| M | -0.67 | 0.52 | -0.76 | | -0.78 | -0.80 | -0.67 | -0.87 | -0.64 | -0.85 | -0.85 | -0.63 | -0.62 | -0.89 | -0.72 | -0.78 | -0.76 | -0.82 | -0.71 | -0.82 | -0.73 | -0.91 |
| N | -0.75 | 0.37 | -0.45 | -1.03 | -0.87 | -0.91 | -0.75 | -0.80 | -0.93 | -0.76 | -0.79 | -0.65 | -0.78 | -0.79 | -0.72 | -0.88 | -0.74 | -0.91 | -0.77 | -0.78 | -0.90 | -0.82 |
| P | -0.77 | 0.36 | -0.81 | -0.92 | -0.75 | -0.77 | -0.88 | -0.73 | -0.78 | -0.81 | -0.69 | -0.94 | -0.90 | -0.88 | -0.85 | -0.80 | -0.78 | -0.88 | -0.91 | -0.70 | -0.82 | -0.85 |
| Q | -0.61 | 0.39 | -0.67 | -0.58 | -0.61 | -0.82 | -0.64 | -0.49 | -0.72 | -0.51 | -0.66 | -0.61 | -0.78 | -0.85 | -0.58 | -0.59 | -0.65 | -0.56 | -0.87 | -0.58 | -0.65 | -0.77 |
| R | -0.54 | 0.36 | -0.56 | -0.74 | -1.01 | -1.00 | -0.64 | -0.66 | -0.41 | -0.52 | -0.36 | -0.54 | -0.55 | -0.62 | -0.26 | -0.56 | -0.57 | -0.46 | -0.52 | -0.54 | -0.58 | -0.65 |
| S | -0.82 | 0.31 | -0.65 | -0.97 | -0.97 | -0.95 | -0.87 | -0.84 | -0.87 | -0.82 | -0.90 | -0.94 | -0.83 | -0.85 | -1.01 | -0.70 | -0.91 | -0.86 | -0.79 | -0.87 | -0.91 | -1.04 |
| T | -0.77 | 0.31 | -0.70 | -0.72 | -0.92 | -1.01 | -0.76 | -0.78 | -0.72 | -0.72 | -0.90 | -0.84 | -0.78 | -0.73 | -0.80 | -0.87 | -0.75 | -0.85 | -0.87 | -0.91 | -0.90 | -0.93 |
| V | -0.80 | 0.27 | -0.65 | -0.78 | -0.84 | -0.77 | -0.77 | -0.89 | -0.79 | -0.89 | -1.09 | -0.92 | -0.83 | -0.74 | -0.91 | -0.78 | -0.92 | -0.84 | -0.93 | -0.80 | -0.86 | -1.14 |
| W | -0.77 | 0.24 | -0.72 | -0.72 | -1.24 | -0.88 | -0.81 | -0.77 | -1.03 | -0.80 | -0.61 | -0.97 | -1.00 | -0.65 | -1.23 | -0.55 | -0.64 | -0.71 | -0.76 | -0.68 | -0.60 | -1.00 |
| Y | -0.73 | 0.44 | -0.70 | -0.80 | -0.99 | -1.06 | -0.61 | -0.73 | -0.66 | -0.69 | -0.70 | -0.74 | -0.66 | -0.84 | -1.14 | -0.64 | -0.67 | -1.03 | -0.83 | -0.92 | -0.63 | -0.83 |
| Avg | | | -0.73 | -0.85 | -0.85 | -0.88 | -0.79 | -0.80 | -0.78 | -0.76 | -0.85 | -0.84 | -0.82 | -0.80 | -0.84 | -0.76 | -0.78 | -0.85 | -0.83 | -0.82 | -0.77 | -0.94 |

Table 5. Predicted interaction energies for non-redundant pairings of peptide residues with target residues (reweighted Autodock model). Residue-level contributions to binding energies (kcal/mol, as estimated by Autodock-based model) for peptide residues of the type shown in the left-most column when the target residue type in closest proximity is that shown in the top row, averages based on 14,591 non-redundant pairings.. "Unb" column contains the energies for peptide residues that have no target residue with any atoms within 4A of any atoms of the peptide residue.  Mean: -0.82. Standard deviation (σ): 0.17.  Blue: more than 2 σ above mean; light blue: between +1σ and +2σ; grey: between -1σ and +1σ; yellow: between -1σ and -2σ; orange: more than 2σ below mean (note energies are negative, lower value connotes higher energy interaction).

As Figure 25 shows, there appears to be a significant relationship between the relative abundance of aromatic (F, W, Y) residues and overall binding energy, both for energies derived from the PopTop model predictions and for actual energies from the PDBBind training set ($r_P$ of 0.63 and 0.42, respectively).

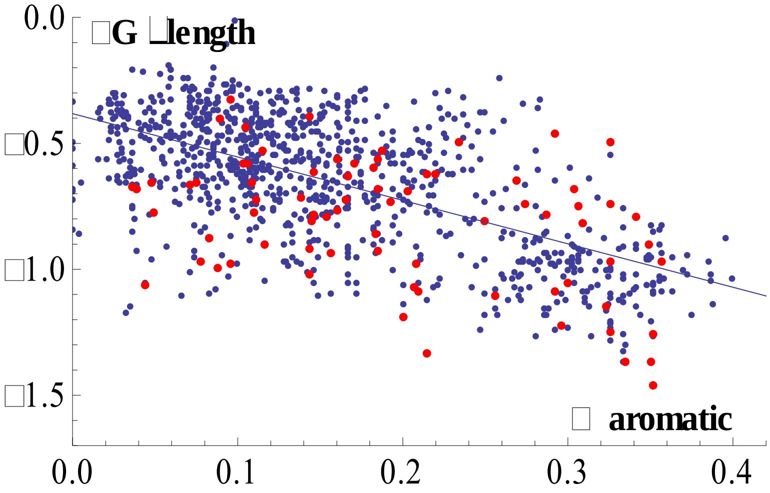

Figure 25. Influence of aromatic and charged residues on estimated binding energies. Binding energies per unit peptide length (SRlen) vs. percent aromatic residues (F, Y, W) shown in SEQRES of peptide plus all I-site residues within 4A of any peptide residue. Blue: energies as estimated by PopTop model. Red: energies from PDBBind database for 75 redundancy groups of interfaces present in PDBBind database.

Figure 26 (blue bars) depicts the relative abundance of the various amino acids among the 'hot spot' peptide residues contributing more than 1.25 kcal/mol to estimated binding energy (the top ten percent of residues in terms of contribution to estimated binding energy), as compared to the overall residue frequencies present in the peptides. The results of a peptide hot spot analysis by London, et al. [185], using a different method (computational alanine scanning) on a 103-interface dataset, are shown by way of comparison (red bars). By the approach used here, the positively charged residues R and K and the aromatics F, Y, and W are the most overrepresented.

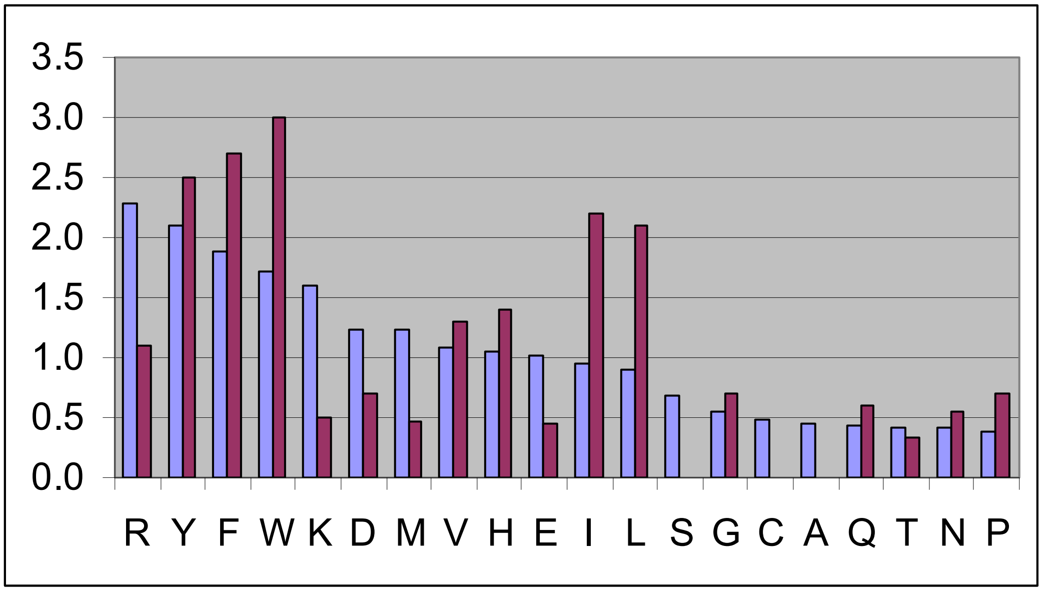

Figure 26. Frequencies of hot spot residues relative to all residues. Blue: ratio of frequencies of residues contributing at least 1.25 kcal/mol of binding energy as estimated by PopTop model to overall residue frequencies in peptides representing 941 redundancy groups (Rgrp). Red: ratio of frequencies of hot spot residues (as determined by 'computational alanine scanning') to overall residue frequencies in peptide-protein interface dataset reported by London et al.

**Abundance and quality of hydrogen bonds, salt bridges, and cation-pi interactions**

Detailed descriptors of hydrogen bond geometry were extracted and included in the Hbonds table in the PPRMint database for each pairing of a hydrogen bond donor atom with a hydrogen bond acceptor atom at a distance of 3.6Ǻ or less. Only pairings of a peptide atom with an I-site atom were included; intra-chain pairings were not analyzed. 30,583 pairings were found, but many have quite poor bond geometry. There are 19,488 H bonds for which the (non-regression weighted) hydrogen bond energy model estimates an energy of -1.0 kcal/mol or better (-8.0 kcal/mol is the optimum for a 'perfect' bond under this model); in 9,503 (48.8%) the peptide is the donor and in 9,985 (51.2 %), the acceptor. 11864 (60.9 %) were bonds to the peptide main chain, and 7624 (39.1 %) to the peptide side chain. These frequencies are in close agreement with those reported by London, et al. [185] (63.0 % and 37.0 %, respectively). The number of H bonds per 100Å$^2$ of interface area was 1.00 ($\sigma$ = 0.55); there were an average of 6.73 ($\sigma$ = 4.28) H bonds per peptide (averages over 941 interface redundancy groups, with interface area taken as the increase in solvent accessible surface area of the target on removal of the peptide). The distribution of bond angles and donor-acceptor distances is depicted in Figure 27(a); the distribution of D-A distances and bond angles in the dataset (see Figure 27(a)) is in close agreement with that found by McDonald, et al. for hydrogen bonds in proteins [223]. The great majority of bonds have donor-acceptor distances close to the mean of 2.90 ($\sigma$ = 0.2Å), and bond angles in the range 140º to 180º. Figure 27(b) shows the distribution of energies as estimated by the (raw, non-regression weighted) hydrogen bond model already described.

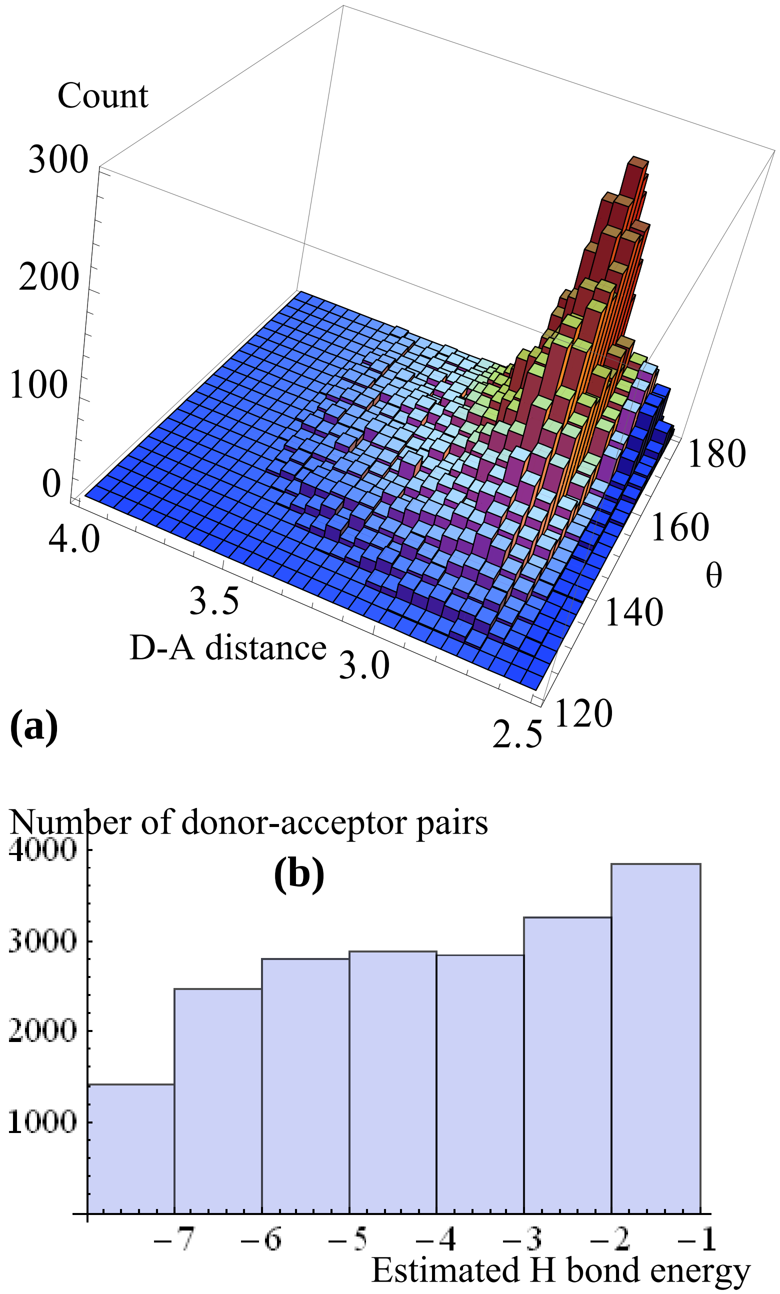

Figure 27. Hydrogen bond geometries. Histograms of 19,496 hydrogen bond donor-acceptor pairs having energy as estimated by raw (non-regression weighted) H bond model ≤ -1.0 kcal/mol, (a) by donor-acceptor distance (Ǻ) and angle made by donor, donor hydrogen, and acceptor (θ, degrees); color indicates estimated energy on scale from -1.0 kcal/mol (dark blue) to -8.0 kcal/mol (dark red) (based on simplified version of H bond energy as function of D_A distance and angle only). (b) by energy as estimated by raw H bond model.

Similarly, the SaltBridges table in the PPRMint database includes records for all pairings of a lysine ζ-N, arginine η-N, or an N-terminal backbone N atom with an aspartate δ-O, glutamate ε-O, or C-terminal carboxyl O atom between the peptide chain and the I-site, at a distance of 5Å or less. Of 7,090 such pairings, 6,049 had energies better than -1.0 kcal/mol as estimated by the (raw, non-regression weighted) salt bridge model. In 3,958 (55.8%) of these pairings, the cation belonged to the peptide; in 3,132 (44.2%), to the I-site. The average number of salt bridges per 100$Å^2$ of interface area was 0.31 (σ = 0.19); there were an average of 2.02 (σ = 1.34) salt bridges per peptide (averages over 941 interface redundancy groups). Table 6 shows the breakdown of salt bridges by type; the most abundant, accounting for 29.6% of all salt bridges, were pairings of arginine or lysine in the peptide with glutamate in the target. Figure 28 compares the distribution of cation-anion separations as measured for the PPRMint dataset with the theoretical salt bridge energy function that was used to compute $E_{SB}$.

The CationPi table of the database encompasses all pairings between peptide and I-site involving phenylalanine, tyrosine, or tryptophan on one side of the interface and a charged amine of lysine, arginine, or the backbone N terminus on the other, if the cation to ring centroid distance is ≤ 8Å. Of 7,564 such pairings, 746 were sufficiently far from optimal to produce estimated energies worse than -1.0 kcal/mol according to the energy model used. Of the remaining 6,818, the peptide contributed the cation in 5,022 cases (73.7%), the aromatic ring in 1,796 (26.3%).

| Cation in peptide | | | Anion in peptide | | |
|---|---|---|---|---|---|
| Type | Count | Percent | Type | Count | Percent |
| K-D | 381 | 5.4 | D-K | 432 | 6.1 |
| R-D | 787 | 11.1 | D-R | 595 | 8.4 |
| K-E | 881 | 12.4 | E-K | 704 | 9.9 |
| R-E | 1216 | 17.2 | E-R | 479 | 6.8 |
| N-D | 135 | 1.9 | D-N | 2 | 0.0 |
| N-E | 489 | 6.9 | E-N | 7 | 0.1 |
| K-O | 34 | 0.5 | O-K | 679 | 9.6 |
| R-O | 20 | 0.3 | O-R | 224 | 3.2 |
| N-O | 15 | 0.2 | O-N | 10 | 0.1 |
| Total | 3958 | 55.8 | Total | 3132 | 44.2 |

Table 6. Salt bridges by type**.** Counts of cation-anion pairings ≤ 5Å across interfaces. Types: K = lysine, R = arginine, D = aspartate, E = glutamate, N = N-terminal backbone N, O = C-terminal carboxyl O. In each pair, peptide residue appears first, followed by I-site residue.

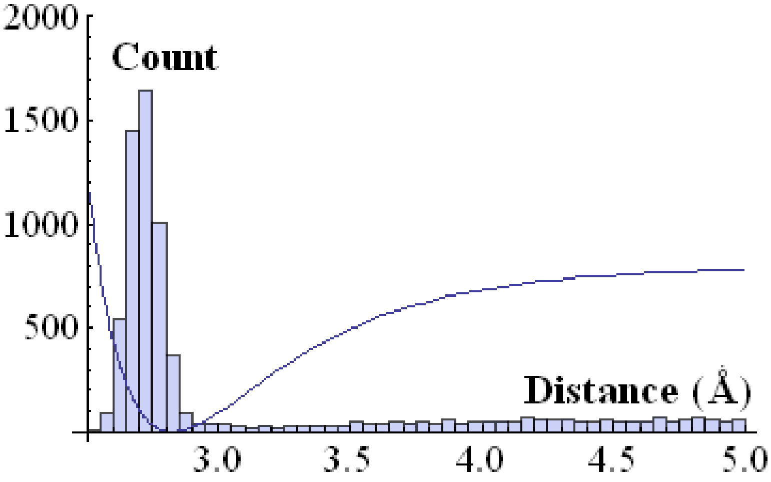

Figure 28. Distribution of salt bridges by cation-anion separation. Histogram of salt bridges by cation-anion distance (Å), with salt bridge model energy function overlaid for comparison (well depth 8 kcal/mol).

The average number of such interactions per $100\text{Å}^2$ of interface area was 0.39 (σ = 0.41); there were an average of 2.53 (σ = 2.81) per peptide (averages over 941 interface redundancy groups). As Table 7 shows, the most abundant pairings, accounting for 37.4% of the total, were of a cation in the peptide with tyrosine in the I-site; the importance of tyrosine in molecular recognition has been noted by others [224]. The distribution of cation to ring centroid distances and angles is shown in Figure 29, and illustrates that a majority of the cation-pi pairings are quite suboptimal.

| Cation in peptide | | | Ring in peptide | | |
|---|---|---|---|---|---|
| Type | Count | Percent | Type | Count | Percent |
| K-F | 294 | 4.3 | F-K | 220 | 3.2 |
| K-W | 258 | 3.8 | W-K | 76 | 1.1 |
| K-Y | 466 | 6.8 | Y-K | 212 | 3.1 |
| R-F | 607 | 8.9 | F-R | 557 | 8.2 |
| R-W | 589 | 8.6 | W-R | 157 | 2.3 |
| R-Y | 834 | 12.2 | Y-R | 565 | 8.3 |
| N-F | 386 | 5.7 | F-N | 1 | 0.0 |
| N-W | 334 | 4.9 | W-N | 4 | 0.1 |
| N-Y | 1254 | 18.4 | Y-N | 4 | 0.1 |
| Total | 5022 | 73.7 | Total | 1796 | 26.3 |

Table 7. Cation-pi interactions by type. Counts of cation-pi pairings with $r \leq 8\text{Å}$ across interfaces. N refers to N-terminal backbone N; in each pair, peptide residue appears first, followed by I-site residue.

### Discussion

A dataset was constructed of 3,924 representative peptide-protein interface structures using PDB data, and a number of structural and geometric characteristics of those interfaces were extracted and quantified. From these, energetic contributions of non-bonded interactions were estimated and a relational database was constructed to facilitate extraction of statistics and testing of hypotheses regarding the contribution of various factors to peptide binding. From this data inferences were drawn that support a tentative set of heuristics to guide the construction of peptide libraries and the selection of candidate peptide sequences for evaluation as ligands, and that may provide useful direction for further inquiry.

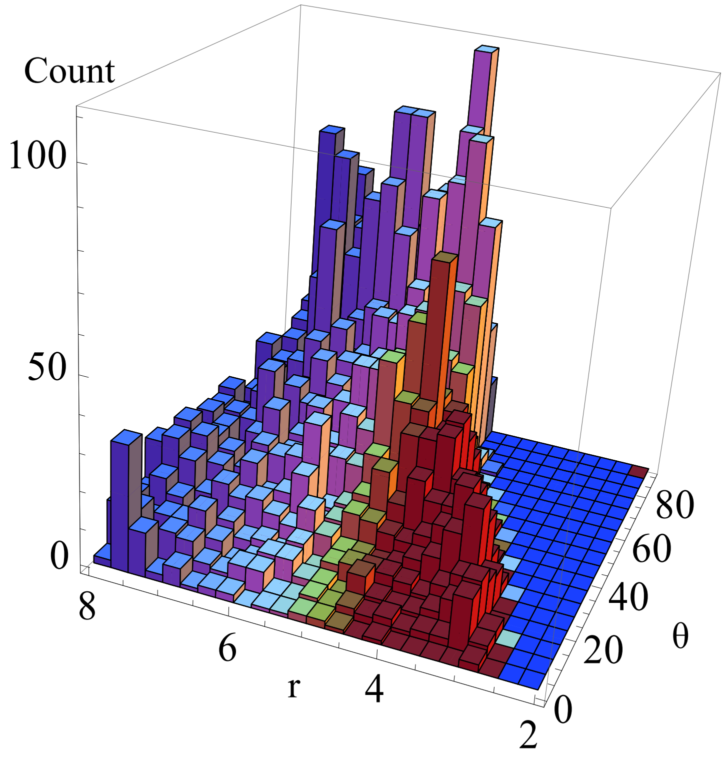

Figure 29. Cation-pi geometries. Histogram of 6,818 pairings of cations with aromatic rings between peptides and I-sites where energy as estimated by model is ≤ -1.0 kcal/mol, by cation to ring centroid distance (r, Å) and angle between

line from cation to ring centroid and line perpendicular to plane of ring (θ, degrees). Color indicates estimated energy on scale from -1.0 kcal/mol (dark blue) to -8.0 kcal/mol (dark red).

### Binding energy prediction model

The choice of energy scoring functions was constrained by the desire to enable evaluation of the relative quality and importance of specific interactions whose contribution to peptide-protein binding were desired to be analyzed in detail -- hydrophobic interactions, hydrogen bonds, salt bridges, cation-pi interactions, van der Waals forces, and electrostatic interactions – and to enable estimates of the contributions of specific residues to the binding energy of the entire interface. The choice was therefore made to base the model primarily on descriptors representing estimates of the actual energies of the interactions of interest, computed using functions describing energies in terms of the specific geometry and characteristics of each individual interaction. It was discovered that a scoring function based on these energy descriptors alone, whether expressed in six separate parameters as in the PopTop model or in three aggregated parameters as output by Autodock, provided only mediocre predictive performance (see Table 1). After experimentation to determine what additional parameters would be most informative, to both models were added three general descriptors of the peptide's size and geometry: radius of gyration (Rg), length in number of residues (Plen), and solvent accessible surface area (ASA), producing considerable improvement in performance. The scoring function using the six separate PopTop energy descriptor terms plus the three geometric descriptors, which is the scoring function by which the energy statistics reported herein were calculated except where otherwise noted, correlated with the PDBBind-derived training set at $r_P$ = 0.72 (ME = 1.07 kcal/mol (0.78 pKD units); on leave-one-out cross-validation, rP = 0.60 and ME = 1.24 kcal/mol ( 0.90 $pK_D$ units)).

There is an extensive literature on binding energy scoring functions, ably reviewed by others [53, 225], all addressing functions primarily or exclusively designed for and trained on protein interactions with small molecules. Wang, et al. [225] benchmarked 14 scoring functions on the full PDBBind 'refined' dataset of (then) 800 protein-ligand complexes [204]. The best performer had $r_P$ = 0.566 and ME = 1.95 kcal/mol (1.42 $pK_D$ units), and the average was $r_P$ of 0.39 and ME of 2.19 kcal/mol (1.59 $pK_D$ units). The average standard error on cross-validation reported for the Autodock 4 energy scoring function [205] was 2.63 kcal/mol (1.91 $pK_D$ units). A scoring function recently reported by Sotrifer et al. [54], using seven descriptors selected as most informative from a 66-descriptor dataset, achieved a $Q^2$ on cross-validation of 0.72 and standard error of estimate of 1.08 $pK_D$ units (approximately corresponding to $r_P$ of 0.85 and $S_{PRESS}$ of 1.49 kcal/mol). Imperfect though the comparison obviously is, the performance of the PopTop scoring function seems reasonably within the range achieved by others, even without taking into account that peptides arguably present a more difficult predictive task owing to the large number of rotatable bonds (which add to the uncertainty regarding the unbound state), and owing to the difficulty of measuring peptide affinities accurately and the consequent likelihood of greater experimental error in the training dataset.

Table 2 shows the fitted descriptor weights for both models, as well as the relative contribution of each descriptor to the computed total interface energies. The large contribution of the peptide length descriptor in both models evidences a strong general correlation between peptide length and affinity. It may be doubted that this correlation holds true for all possible peptide sequences, keeping in mind the selection bias inherent in working from a training set containing only peptides actually bound in interfaces. Nevertheless, for peptides known to bind at some measurable affinity, it does not seem unreasonable to suppose that each additional residue makes some contribution to affinity, on average, assuming it is able to make productive contact with the protein surface. Viewed in this way, the peptide length descriptor may be regarded as representing a rough aggregate of the average per-residue contributions of the various forces, and the weighting of the descriptors of the actual forces as in effect fine tuning the contribution of each as it differs from the aggregate embodied in the weighted length descriptor. (Indeed, in most if not all parameterized energy scoring functions, there is considerable overlap between descriptors – for example, a salt bridge term and an electrostatic term can hardly be said to be perfectly orthogonal.) Accordingly, the mean contribution of each of the PopTop energy-related descriptors is relatively small, and each has a range and variance such that the contribution can be either positive or negative for a given interface. The considerable improvement in predictive accuracy obtained by including the size-related terms suggests a likelihood that the energetics of these peptide interfaces are not particularly well captured by descriptors that focus only on the details of the atom-level non-bonded interactions. Several possible reasons suggest themselves. First, as discussed in greater detail below, analysis of those interfaces in the PPRMint dataset that are present in multiple redundant versions strongly suggests that peptide-protein interfaces are much more mobile and dynamic than ligand-protein interactions are usually assumed to be, with many transient non-bonded interactions, so the non-bonded interaction descriptors are arguably aiming at a moving target. Second, with respect to the hydrogen bond descriptor and to a lesser degree the salt bridge descriptor, it must be kept in mind that any available donors or acceptors in either the peptide or the I-site are likely, in the unbound state, to be hydrogen bonded with water, so that a hydrogen bond in the peptide-protein interface does not necessarily reflect a large change in energy (and, if the bond geometry is poor, may even be energetically inferior to the solvated state). Third, for a molecule having the flexibility of a typical peptide, it is not surprising that most of the contribution of van der Waals forces could be encompassed in a gross per-residue aggregate, since it would be expected that the peptide would arrange itself in such a way as to avoid greatly suboptimal inter-atomic distances.

### Interface composition and redundancy

Any peptide-protein interface structure dataset, regardless of how selected, is inherently very far from an unbiased sampling of peptide-protein affinity space.

The PDB contains only those complexes that are either amenable to NMR analysis or that can be crystallized, and then only those that involve molecules of sufficient interest to justify the effort and expense. Nearly all peptides in the PDB are of biological origin; there is essentially no sampling of random or artificial sequences.  Obviously, therefore, no claim is made that the PPRMint dataset is representative of peptides in general; it does, however, represent a kind of 'existence proof' of what outcomes can reasonably be expected from those combinations of peptide and interface characteristics that are adequately represented in the sample.

It is common in assembling structural datasets to make careful selection of the structures presumed most accurate, typically by including only X-ray structures of resolution less than some fairly strict cutoff, and to avoid inclusion of structures presumed redundant, typically by rejecting any structures having sequence and/or structural similarity to any of the structures retained. The goal of all such analyses is to extract the most accurate information possible from raw data that is noisy and represents a sparse and far from random sampling of the search space. Careful exclusion of all but the (presumed) highest quality data is one strategy for doing so, but it is not clear that the presumed improvement in accuracy necessarily compensates for the resulting reduction in sample population.  Particularly in the context of peptide-protein interfaces, rejection of all but the very highest-resolution X-ray structures has several negative consequences.  It greatly reduces the diversity of a sample population that is already badly biased by virtue of including only structures that happen to have been studied and solved. It ignores all of the inaccuracies arising from causes other than poor resolution, such as guesswork in model fitting. Most importantly from the standpoint of the present analysis, it implicitly assumes that a single structure can adequately represent an interface that, according to the evidence presented in Chapter 2, may be quite dynamic and mobile. A different approach was therefore chosen, and the decision was made to include in the PPRMint dataset all peptide-protein pairings available in the PDB at the time the sample was taken, if they could fairly be characterized as interactions of an 8- to 32-mer peptide with a surface of a protein. NMR structures as well as X-ray structures were included, regardless of resolution and redundancy.  The approach chosen opens the possibility of employing methods analogous to the use of signal averaging to measure a weak, noisy electrical signal, an approach that produces results far superior to those that would be obtained by attempting to identify and select the single "best" sample from a sample set.

A further reservation regarding the common practice of relying on sequence similarity cutoffs to eliminate redundant data is that whether or not two data points are redundant depends on what information is sought. The inclusive approach allows grouping of data according to similarity criteria tailored to each property being analyzed, and at a level corresponding to the property of interest – extracting groups of similar interfaces for interface level properties, groups of similar residue pairs for properties at the level of residue interactions, etc. Since the detailed characteristics of each interface are exposed in the relational database, it requires only a simple SQL query to extract subsets or groupings based on any desired combination of criteria.

### Implications for peptide ligand design

Based on the results already described, several observations may be offered that may provide useful insight regarding the design of peptide libraries and the selection of peptide leads.

*1. A practicable optimized affinity limit is about 10 nM kD.* As Figure 21(b) shows, for peptides in the 8 to 20 residue size range, the energy distribution rarely extends below approximately -11 kcal/mol, corresponding to $K_D$ of approximately 10 nM.

*2. Peptide length is relatively unimportant.* As also appears from Figure 21(b), the distributions of estimated binding energy are similar for all lengths from 8 to about 20, with a few outliers in the -13 kcal/mol range for lengths above 12. The data suggests that somewhat better energies can be obtained for lengths above 20, but since the average marginal gain per residue is then only about 0.4 kcal/mol on average, any gain may be illusory due to offset by the increase in entropic penalty. Given the apparent lack of dependence on length, an effective strategy may be to screen on libraries of (say) 20-mers, with the expectation that doing so increases the probability of the peptide having some sub-region capable of binding the target as compared to shorter peptides, and that any 'hits' might then be optimized down to a shorter length.

*3. Peptides prefer to bind in 'arroyos'.* It is well understood that polypeptides find it energetically advantageous to fold in such a way as to hide hydrophobic moieties from the solvent, so it is not surprising that interface configurations that remove more of the peptide surface from exposure to solvent would be favored, at least to the extent that doing so tends to bury hydrophobic regions more so than hydrophilic ones. It is also logical that moieties in contact are usually contributing energy to the interface, since otherwise they would presumably tend not to remain in contact. The data shows a clear energy gain on average from increased burial of peptide surface (Figure 22(a)), or, expressed in another way, from larger numbers of I-site residues in contact with the peptide (Figure 22(b)). Analogously, London, et al. concluded, using other methods, that peptides "tend to bind in the largest pockets available on the protein surface" [185].

*4. Aromatic residues in the interface enhance binding.* The data show tryptophan and tyrosine fully three-fold overrepresented in I-sites, and phenylalanine 1.5-fold overrepresented (see Figure 16).  At least over the range spanned by the dataset, the interfaces with better estimated energies tend to be characterized by relatively higher percentages of aromatic residues (Figure 25). In the 'hot spot' residues making the greatest contribution to estimated binding energy, the aromatics are the most overrepresented (2.5-fold, 2.7-fold, and 3-fold for tyrosine, phenylalanine, and tryptophan, respectively, see Figure 26). The aromatics were also the most abundant ‘hot spot’ residues in the 103-interface dataset analyzed by London et al. using different methods [185].

*5. Peptides should be relatively unstructured (but not too much).* In theory, an overly flexible peptide should bind poorly, other factors being equal, because of the larger decrease in conformational entropy on binding. On the other hand, a peptide that maintains an overly rigid structure should be less able to adapt to the target surface so as to find an optimal fit. One measure of “foldedness” relates to the tendency of polymers to exhibit random flight behavior under θ conditions and expand or collapse as conditions deviate from θ [51]. A substantial majority of the peptides in the dataset have end-to-end lengths greater than the theoretical random flight lengths, suggesting a relatively un-collapsed state. See Figure 18. It may be noted, however, that a number of the interfaces in the shorter length ranges having the best estimated binding energies do have end-to-end lengths somewhat smaller than their theoretical random flight lengths, consistent with the hypothesis that peptides short enough to comprise a single “blob” (see Chapter 2) may have relatively stable conformations, which may lead to excellent binding if that conformation happens to be highly complementary to a binding site on the target. It may be hypothesized that the optimal state for peptides in relatively small libraries may be somewhat extended as compared to the θ state, so as to provide both reasonable diversity and adequate affinity. As Figure 23 shows, the peptides with the best per-residue energetics appear to cluster around an end-to-end length to random flight length ratio of about 2. The distribution of secondary structure tendencies arguably supports the view that excessive flexibility is undesirable: for the peptides that have the best per-residue estimated energies, the longest contiguous stretch of residues in the extended-β compartment of Ramachandran space comprises about 20% of the peptide length (see Figure 19).

*6. Salt bridges enhance binding.* Pairings of arginine or lysine with aspartate or glutamate were by far the highest frequency pairings in the dataset (Table 3): of the 11,871 peptide residues in contact with the I-site, for fully 1,457 (12.3%) of these, the peptide residue was one of the charged residues, and the closest I-site contact was one of the oppositely charged residues.  In the great majority of the charged residue pairings, the atomic spacing between the charged atoms was close to ideal, as shown in Figure 28. In terms of estimated energy contribution, pairings in which the peptide residue is the cation are favored over those in which the peptide residue is the anion, with arginine-aspartate, arginine-glutamate, and lysine-aspartate pairings all estimated to contribute energy at least three standard deviations above the average for all pairings.  Arginine and lysine are also among the most overrepresented types among 'hot spot' residues in peptides (those contributing at least -1.25 kcal/mol to the estimated binding energy, see Figure 26).  It seems reasonable to infer from this data that the salt bridges, an average of approximately two per peptide, provide relatively fixed 'anchors' tethering the peptide to the I-site and stabilizing the overall positioning of the peptide in the interface.

*7. Hydrogen bonds in interfaces are relatively unimportant.* Hydrogen bonds were considerably less abundant (1.00 bonds per 100Å$^2$ of interface area (σ = 0.55)) than reported by London, *et al.*, for their database of 103 peptide-protein interfaces (1.6 H bonds per 100Å$^2$), but in agreement with the H bond density of 1.0 per 100Å$^2$ in protein-protein interfaces found by Xu et al. [188]. The net regression-weighted contribution of the (average) 6.73 hydrogen bonds per interface ranged from -1.49 to + 3.15 kcal/mol, with an average of +0.83 (σ = 0.52); in effect, on average, hydrogen bonds tended to worsen estimated binding energy. As previously noted, this may be, in part, an artifact arising from the non-orthogonality of the hydrogen bond energy term and the length and area terms of the energy function. More importantly, however, it appears that many of the hydrogen bond donors and acceptors present in the bound state find themselves in quite suboptimal geometries where their energetic contributions are far from ideal (Figure 27(b)), providing little improvement, if any, over the solvated state in which they are hydrogen bonded to water. The energy of hydrogen bond donor-acceptor pairings is quite sensitive to small changes in geometry, and because they are relatively abundant, with many suboptimal pairings present, any motion of the peptide appears likely to worsen the geometry of some pairings while improving that of others. For all these reasons, it may be hypothesized that in designing peptides for optimal affinity to proteins, there is little to be gained by attempting to engineer hydrogen bonds into the interface.

*8. Binding can likely be improved by increasing the opportunities for cation-pi interactions.* Given the already noted strong overrepresentation of aromatic residues in I-sites and their correlation with higher energies, and given the high abundance of arginine and lysine in the peptides in the dataset, it is not surprising that cation-pi interactions appear to play an important role in these interfaces. Cation-pi interactions are of particular interest because of the potential for energies approximately double or more that of an optimally positioned hydrogen bond, with considerably less dependence on exact positioning [213]. Pairings of cations with aromatic rings with energies better than -1.0 kcal/mol according to the raw (non-regression weighted) cation-pi model used were somewhat more abundant than salt bridges in the interfaces, averaging about 2.5 per interface. In the majority (73.7%), the cation belonged to the peptide and the aromatic residue to the I-site. According to the model employed, the energy of a cation-pi interaction is optimal when the distance from the positively charged atom to the ring centroid is about 2.8Ǻ, and the cation is positioned on the perpendicular axis of the aromatic ring. As Figure 29 shows, the pairings of cations with nearby aromatic rings spanned a wide range of geometries, most quite far from ideal. It may be hypothesized that the significance of the distribution shown in Figure 29 is that peptide-protein interfaces are likely to contain cation-pi pairings that make energy contributions that are not insignificant even when the geometry is suboptimal, and, more importantly, have the potential to improve as the peptide shifts in position.

*9. It is not essential for the peptide and the I-site to be oppositely charged.* An unexpected finding was that absolute charge complementarity at the interface level does not appear to be of great importance. As shown in Figure 15, although charge-complementary interfaces are more common in the dataset than interfaces in which the estimated net charge of the peptide is of the same sign as that of the I-site, there are many counterexamples. Charge complementarity of charged residues individually is, as expected, nearly absolute; it may be that interface-level charge complementarity would tend to correspond to a relatively non-specific tendency to bind any oppositely charged protein surface, and that in interfaces in which the peptide has been selected for specificity, a complementary pattern of mixed charges would be expected. In support of this hypothesis, it may be noted that many of the peptides in the dataset contained mixed charges (see Figure 14 for distribution).

*10. Peptides in interfaces are not rigid or immobile.* The data reported here tends to support the hypothesis, advanced by others both for protein-protein interfaces [226] and for peptide-protein interfaces [185], that a few "hot spot" residues contribute disproportionately to the binding energy in a peptide-protein interface (see Figure 26). It may be suggested that the hot spot hypothesis, taken to its logical conclusion in the context of peptide-protein binding, implies a model in which the peptide is viewed as a relatively dynamic entity, in which a few hot spot interactions anchor the peptide to the protein surface at relatively fixed loci, and the peptide regions between hot spots make relatively modest or negligible contributions to ΔG and are to some extent mobile. Although the obvious limitations of x-ray and NMR structures for drawing conclusions about ligand mobility must be acknowledged, analysis of groups of redundant interfaces (Figure 20) does appear to indicate that peptide binding is considerably more dynamic than a "lock and key" model would suggest. In most cases, the bound conformations of the peptides in the dataset appear unlikely to represent stable folded shapes (Figure 18 and Figure 19). The relative overrepresentation of positively charged residues and aromatic residues in hot spots implies that the hot spot interactions are likely to favor salt bridges and cation-pi interactions. Both of these have the potential for relatively high energy contributions, and both can tolerate moderate changes in position – particularly rotations that do not greatly affect separation distance. Clearly, a very few reasonably configured salt bridges and/or cation-pi interactions are sufficient to contribute energy on the order required to attain the affinities observed. It may be hypothesized that the energy contributed by the non-hot spot regions of the peptide may be attributable to the effect of attractions such as hydrophobic forces, electrostatics, and van der Waals attractions, that combine to produce a very modest and relatively non-position specific attraction, and to moieties that participate in weak hydrogen bonds or poorly configured cation-pi interactions that are easily traded for others as the position on the protein surface changes.

*12. Library peptides should be overrepresented in R, K, W, Y, F, I, L, D, E, and P.* Based on abundances in hot spot residues and higher affinity interfaces, and on the higher estimated energies for suitable pairings involving these residues, it appears that the residues most likely to contribute disproportionately to the affinity of a peptide ligand for a randomly chosen protein target are the aromatics W, Y, and F; the positively charged residues R and K, and the hydrophobic residues I and L. D, E, or both should also be included to provide an oppositely charged partner for pairing with positive charges on the protein surface. One other possibly desirable inclusion is proline. The data shows a quite high frequency of P-W and P-Y pairings, and although the trained energy model does not predict an unusually high energy attributable to them, it may be suspected that these pairings are over-represented in interfaces for a reason. Although there are a few poly-proline subsequences in the dataset, contiguous peptide subsequences of 3 or more prolines account for only 11% of the non-redundant P-W and P-Y pairings, so the reason does not appear to be bias on account of poly-prolines. It may be speculated, based in part on observations from molecular dynamics simulations of peptides, that peptides may tend to benefit from the well known tendency of proline residues to introduce bends in the chain, facilitating the tendency of the peptide to non-specifically fold over against itself in solution and minimize hydrophobic surface, thereby reducing its conformational diversity in solution and consequently the entropic penalty on binding, while preserving sufficient mobility to allow adaptation to the protein surface.

From the foregoing observations, the tentative outlines of a design strategy may be seen to emerge. The first step would entail an analysis of the surface of the protein target to identify contiguous surface regions that are enriched in aromatics and charges and that are located in a depression or ‘arroyo’. Within those regions, residues suitable for anchoring the peptide hot spot residues would be identified. A candidate peptide sequence having a length in the 12 to 20 residue range would be defined by first selecting aromatics and charged residues suitable for the anchor sites on the protein surface – R or K in juxtaposition to protein surface aromatics or negative charged residues, and aromatics or D or E residues opposite positively charged protein surface residues. “Spacer” residues would then be filled in between the hot spot residues, taking into account the distances between adjacent intended anchor loci. It may be possible to fine tune the specificity / affinity tradeoff by adjusting the number and positions of I and L residues in the spacers, to obtain more or less favorable hydrophobicity matching. And, if it is possible to include a proline residue in the middle region of the peptide while maintaining reasonable compatibility with the inter-anchor distances notwithstanding the resulting proline-induced kink, doing so may be worthwhile.

A similar strategy might be employed to prioritize leads obtained from peptide microarray selection: the candidate sequences might be analyzed for conformity to the criteria just described, in order to select a smaller number for testing and optimization.

It must be emphasized that the design process just described is not envisioned as a substitute for random library screening, which remains essential. If anything, the results reported here imply a binding mechanism much more dependent on unpredictable stochastic processes than is thought to be the case for rigid, lock and key models. The intent is to infer ways in which the random screening might be made more efficient by biasing the random library away from sequences that are less likely to produce useful leads.

### Conclusions

Detailed geometric, energetic, and other descriptors have been computed for a dataset of 3,924 minimized peptide-protein interfaces extracted from the PDB, and compiled in a publicly accessible relational database (see http://www.innovationsinmedicine.org/pprmint/pprmint.mdb). Using this data, statistics have been extracted on various properties of the peptides and their binding sites that can be estimated and/or designed for, with a view to obtaining heuristics of potential use in designing and optimizing peptide libraries and in prioritizing leads. The data suggest that it should be feasible to improve the efficiency of peptide ligand selection using libraries of peptides in the 12 to 20 residue range, enriched in charged and aromatic residues, and that it should generally be possible after selection and optimization to obtain peptide ligands having affinities as high as approximately 10 nm.

# CHAPTER 4: PREDICTION OF PROTEIN-PEPTIDE INTERACTIONS BY VIRTUAL SCANNING PROBE MAPPING

## Abstract

Peptides represent an increasingly important class of ligands for use in therapeutics and diagnostics. Peptide-protein binding depends in significant part upon protein surface physical and chemical characteristics that can be estimated and mapped computationally.

Peptide binding loci were predicted for eight randomly selected peptide-protein complexes, using Virtual Scanning Probe Mapping (VSPM), a new strategy for spatially mapping the interactive properties of a macromolecular surface of known composition and geometry by systematically interrogating the surface computationally with molecular probe entities containing moieties representative of the interactions of interest. As applied to a test set of eight PDB structures involving peptides from 8 to 16 residues in length, in complex with proteins and for which solved structures of both the bound complexes and the unbound proteins are available, the set of protein surface residues predicted to comprise the peptide binding site included at least 25% or more of the binding site residues in seven of the eight cases, and at least 50% in four of the eight cases, for a mean true positive rate of 45% and false positive rate of 9% for residue-level predictions based on the bound form protein structure. For predictions based on the unbound protein structures, at least 24% of the correct

binding site residues were predicted in six of the eight cases, and at least 49% in four of the eight.

The VSPM method predicts peptide binding loci with an accuracy that compares favorably to other available computational strategies, makes correct predictions in some cases where other methods fail, and, being entirely physics-based, can be applied to arbitrary structures that include non-natural residues. VSPM furnishes a general strategy of potential use in applications where the spatial distribution of the interactive characteristics of a macromolecular surface is of interest.

### Background

In keeping with the considerable and growing interest in the discovery of peptide ligands for use in therapeutic, diagnostic, and other applications, a major research focus in the Center for Innovations in Medicine has been the development of methods for engineering multivalent peptide-based ligands capable of participating in specific, high-affinity interactions with protein or other macromolecular targets [29-32]. This has involved the use of a large library (currently about 10,000) of random sequence 20-mer peptides to investigate the affinity behavior of peptides in interactions with a variety of biological analytes, typically in a spotted microarray format or by SPR. From these experiments, lead peptides are selected and optimized. In addition to its obvious relevance to the effect of ligand binding on target function, information identifying the region(s) preferred by each peptide on the surface of the target protein is of particular interest in the context of designing multivalent peptide-based ligands, where the selected peptides should bind at distinct loci and where the geometry and dimensions of the multivalent ligand should be consistent with the required spacing between the binding loci of the component peptides. Here a general computational method is described for mapping the spatial distribution of protein surface interactive properties, and an evaluation is made of its performance in predicting the peptide binding loci in a test set of eight randomly chosen protein-peptide complexes for which PDB structures of both the bound complex and the unbound protein are available.

The method proceeds by scanning the target protein surface in a manner analogous to scanning probe microscopy, but using "virtual" probes selected for their ability to capture particular surface properties of interest. The interactive tendencies of these probes at various points on the surface are determined and mapped computationally. The goal is not to ask precisely *where* a peptide binds, a question that impliedly assumes a single binding locus and arguably ignores the probabilistic nature of peptide binding kinetics. Instead, the objective is to estimate the spatial distribution of protein surface binding propensities with respect to each of the interactive moieties exposed by the peptide, keeping in mind that residues present in flexible peptide ligands can likely interact, with varying energies and dwell times, at multiple loci. As reported here, by interrogating the surface of a protein for which a solved structure is available, using multiple small probe entities representative of the various residue combinations present in a peptide of interest, constructing spatial maps of the energetics of the interactions of each probe with the protein surface, and aggregating and mapping the information obtained thereby, it is possible to make surprisingly accurate predictions of the highest probability binding sites and, in some cases, the approximate structural arrangement of the peptide.

### Peptide binding site prediction

Available experimental strategies for determining peptide binding loci on proteins are not practicable on any significant scale. Methods such as x-ray crystallography of bound peptide-protein complexes, SAR by NMR [227], and deuterium exchange mapping [228, 229], are costly and time-consuming, often require artificial conditions that may not represent accurately the interaction of interest, and are likely to fail or produce misleading results when applied to relatively unstructured regions or entities. Some insight can be obtained by cross-linking bound peptides to the protein surface and locating the sequence positions of the cross-linked residues by mass spectroscopy [229-232], but such experiments depend on the presence of cross-linkable moieties at or near the peptide binding site, are technically demanding, and (typically) locate only the N-terminus of the peptide and that only within a radius commensurate with the size of the cross-linker (typically on the order of 1 nm). See Chapter 5.

An alternative is to attempt to model the protein-peptide interaction using computational tools. Established methods include computational docking [233, 234]; bioinformatic prediction based on sequence or motif similarity with known complexes [235]; and identification of potential binding hot spots based on surface chemical properties [214].

Computational docking has proved impracticable in the context of peptides and proteins in the size ranges of interest, for several reasons. First, the size of the search space increases exponentially with the number of conformational degrees of freedom in the ligand, so computational docking typically works poorly when applied to peptides more than a very few residues in size. Although impressive progress has been made in extending the ability of docking search to cope with conformational flexibility [236, 237], the number of torsional degrees of freedom present in a typical 8- to 20-residue peptide exceeds current capabilities. Second, if the flexibility of protein surface residues is also taken into account, the complexity of the search multiplies by additional orders of magnitude, but if this is not done, the unbound surface being explored may differ considerably from the actual surface shape present in the bound complex [238, 239]. Third, unlike typical small molecule docking, where the binding pocket or other region being targeted is known in advance, the large size of the peptide in comparison to the target surface and the relative weakness of the interactions between its residues and the target make it necessary to search the entire surface (referred to as "blind" docking), greatly increasing the size of the search space [234]. Although other analytical approaches may be employed that significantly improve efficiency over that of blind docking by pre-identifying the most likely binding sites [240], it is unclear that doing so would greatly reduce the scope of the docking search for the peptides here of interest, which are in any case often so large that a genetic algorithm search over a pre-computed force field grid (such as that employed by AutoDock [205, 206]) would require a very large grid box merely to ensure that the permuted conformations generated by the genetic algorithm would remain within the grid boundaries.

Similarity-based bioinformatic methods are limited by the need for training data on which to base predictions. Structural data on peptide-protein bound complexes is available for only a relatively small number of interacting pairs, so it is unlikely that existing databases will contain structural relationships that are directly comparable to those present in a given unknown interaction, even where the unknown interaction is one that occurs naturally in biology, and much less for interactions involving peptides that have sequences and/or chemical properties not found in nature. Others have had some success in overcoming this lack of training data by sampling structural relationships present in folded proteins [235], but this strategy again presupposes naturally occurring relationships, and in any case often fails when applied to particular complexes.

Several methods have been reported for assessing the more general question of which residues or atoms on a protein surface are most likely to be involved in interactions with other polypeptide entities, based on factors such as evolutionary conservation, statistical analysis or machine learning based on databases of known structures [241], and identification of surface characteristics such as hydrophobicity [214]. Methods of this kind contribute important information, but they suffer from several inherent limitations, particularly when applied to the analysis of the binding propensities of random sequence oligopeptides of arbitrary composition. Random peptides present no evolutionary relationships to rely on; there is no database of evolutionarily or functionally similar complexes to use as a training set. Random sequences seem likely to inhabit regions of conformation space that are uncharacteristic of naturally occurring entities, raising the probability that they will also present combinations of chemical properties that differ from those typical of biological sequences. Existing bioinformatic methods are, in principle, not readily extensible to peptides containing non-natural residues, so their utility is narrowly limited in terms of future expansion of the repertoire of components available for engineering desired properties (a significant drawback given the incentive to make peptide therapeutics resistant to enzyme degradation). Finally, none of the computational methods reviewed (including that reported here) make correct predictions in all cases, or with more than approximate accuracy, so the addition of physics-based strategies to the repertoire of analytical tools should provide another useful informational input.

### Methods

To apply the VSPM strategy to the peptide binding site identification problem, software was written that leverages the ability of the AutoDock 4 suite of computational docking software [205, 206] to estimate the energy of interaction between an arbitrarily defined probe ligand and an arbitrary point on a molecular surface. This strategy extends the docking paradigm beyond the usual focus on a single docking site, and instead asks how the binding energy landscape of the entire surface appears from the standpoint of the probe ligand. In principle, a spatial map could be made of the estimated ΔG of interaction by exhaustively computing energies at each point on the surface at some suitable interval and for multiple orientations and conformations of the probe. The current implementation instead constructs the map by sampling, and expresses the result in terms of the (Boltzmann weighted) likelihood of the probe binding at each surface point. In effect, within the limitations of AutoDock's search algorithm, it simulates an ensemble consisting of a large number of random encounters between the probe and the protein.

The protein surface is interrogated using a set of 'probe' entities that are, taken together, representative of the composition of the peptide ligand of interest, but which are individually small enough for reasonably efficient computational docking search. The protein surface is divided into overlapping regions that cover the entire surface of the target protein. Via multiple parallel runs of AutoDock, a sample population of docked solutions (each referred to as a docking "pose") is generated for each probe entity in each region. After appropriate reweighting for estimated energies and to adjust for various factors that bias the search, the target surface is mapped according to the frequency of samples near each surface atom, resulting in a score for each surface atom that represents the relative likelihood that the atom is part of a region having a propensity to bind the selected probe(s).

The discussion to follow summarizes the steps performed by the software implementation of the VSPM method as applied to the peptide binding site prediction problem. This is embodied in a suite of Python programs that manage the preparation of the inputs (in part via callouts to suitable external utilities); set up, schedule, and reassemble the output from multiple parallel docking runs on the ASU high performance computing cluster, of multiple probes in multiple overlapping target protein surface regions; extract and analyze the outputs, reweight them, and map them to the protein surface; generate statistical comparisons between the predicted and actual binding sites; and, finally, search for combinations of samples that satisfy geometric constraints derived from peptide sequence information.

### Generation of probe entities

A set of probe entities is first defined that is collectively representative of the composition of the peptide of interest. To facilitate the docking search, the probes are limited to structures having no more than approximately ten rotatable bonds. Obvious candidates are otherwise relatively inert entities in which a moiety exposed by the peptide of interest is embedded; and small subsequences from the peptide itself. To arrive at a practicable probe size, blind docking experiments were performed with short peptide sequences in size ranges from two to seven residues. It was found that trimers appeared to represent the best compromise for capturing as much of the specific chemical character of the peptide as possible while keeping the number of torsional degrees of freedom within practicable bounds. For the range of docking search parameters used, AutoDock is able to dock peptide trimers and produce docking poses most of which are substantially in contact with the protein surface, while tetramer and larger peptide subsequences tend often to result in docking poses that are only partially in contact with the protein surface. The results reported here are based on two probe types: (1) probes consisting of one residue present in the peptide, flanked by a single alanine residue in each direction; and (2) probes consisting of the set of all (fully overlapping) trimers generated from the sequence of the peptide of interest. The software generates these probes in automated fashion from the input peptide sequence, producing starting structures for each probe in PDB format using the tleap utility of the Amber 9 molecular dynamics suite [131]. It then converts these to the pdbqt format required by AutoDock, using the prepligand4.py utility supplied with the AutoDock 4 release, accepting the default choice of rotatable bonds.

### Preparation of the target protein for blind docking

AutoDock performs its docking search by attempting to optimize the position and conformation of the ligand within a three-dimensional rectangular box in which force field values have been pre-computed at each point in a grid lattice within the box (and can therefore be efficiently computed at any arbitrary point within the box by interpolation). The VSPM software automates the allocation of the space surrounding the target protein to three-dimensional rectangular grid boxes that overlap in all dimensions by an average of 50 percent, and that together provide multiply overlapping coverage of the entire protein surface. For the structures presented here, the number of grid boxes required ranged from 1 to 24, depending on the size of the target protein, and using the AutoDock default grid lattice spacing of 0.375Ǻ. Finer grid lattice spacings did not appear to improve the overall results significantly, but increased the number of grid boxes required, which in turn increases the computing time required and exacerbates the bias arising from grid box edges discussed below. After computing the accessible surface area of each target protein atom using the utility MSMS [242] with a probe ball radius of 1.4Ǻ, and converting the target protein PDB structure to the required pdbqt format using the AutoDock utility prepreceptor4.py, the software sets up and queues parallel jobs for computation of the grid force field matrices by AutoGrid for each grid box.

### Docking search

After the grid maps have been generated by AutoGrid and the probe structures have been created, sets of input files for AutoDock are created, one set for each probe in each grid box, and these are run in parallel on the ASU high performance computing cluster to produce 25 individual solutions, or "docking poses", for each probe in each grid, with a maximum of 2,500,000 energy evaluations and 27,000 generations of the genetic algorithm per solution. These poses, together with their docked energies as estimated by AutoDock, are parsed from the AutoDock output and converted to PDB format, resulting in a large number (25 times the number of probe entities times the number of grid boxes) of samples representing possible interactions between the probe moieties representing the peptide and the residues exposed on the surface of the target protein. Except as noted, AutoDock default parameters were used.

### Mapping of target surface loci affected by sampled docking poses

The target surface is arbitrarily defined as the set of atoms of the target protein having an accessible surface area of at least 1Å$^2$ as computed by MSMS using a 1.4Å probe ball radius. The binding site of each sampled docking pose is determined by identifying the set of target surface atoms each of which is within a threshold distance of 4Å from at least one atom of the docked probe. The goal is to estimate the extent to which the number of probes that dock to a given target surface locus exceeds the number that would be expected at that locus if the sampled docking poses were distributed randomly. This is computed in the form of a likelihood ratio, in which the numerator represents the actual count of the number of times a target surface atom is part of a probe binding site, and the denominator represents the number of “hits” the same target atom would be expected to receive under the null hypothesis in which probe docking loci are distributed randomly over the protein surface. This computation requires that the expected random hit counts be adjusted to remove several sources of obvious bias, and that the numerator be adjusted so as to weight the probe binding pose solutions by energy. The resulting likelihood ratios are mapped onto the target protein surface by generating scripts for visualization using Pymol 2.4 [243].

Predicted binding sites are identified as the set of all target residues containing at least one atom having a likelihood ratio above a specified cutoff.

### Reweighting of expected hit counts to eliminate sources of bias

A goal of the model is to measure the extent to which the probe docking poses in the sample set exhibit a preference for particular loci on the target surface. The subdivision of the target surface into grid regions for docking potentially biases the resulting sample population in at least three ways:

(1) The aggregate target protein surface area contained within each grid box varies according to the size of the grid box and the topology of the target surface. Since a fixed number of samples are collected for each probe in each grid box, a target surface atom within a grid box that encompasses a larger proportion of the target surface is less likely to be part of a given binding locus than an atom in a grid box encompassing a smaller proportion of the target surface, other factors being equal. This effect is nonlinear with respect to the amount of target surface area enclosed: for a grid box enclosing a small target surface area approximately equal to the surface area occupied by a single docking pose, every docking pose would affect every target surface atom within the grid box, so the *a priori* probability of a given target surface atom being affected by a hypothetical docking would be unity. (Obviously, such grid layouts are to be avoided; the example is offered merely to illustrate the nonlinearity.)

(2) A second source of bias arises from grid overlap. The VSPM software attempts to lay out grid box arrangements with substantial overlap, so as to reduce

the bias arising from edge effects discussed below. Target surface atoms that “belong” to more than one grid box obviously have a correspondingly increased likelihood of being counted as part of a probe binding site.

(3) Additional bias arises from AutoDock’s tendency to disfavor docking sites near the edges of grid boxes. AutoDock uses a genetic algorithm-based search to look for ligand (probe) positions that maximize the energy as computed using the force field grid maps pre-computed by AutoGrid. Poses that result in all or part of the probe being outside the geometric confines of the rectangular grid box are disallowed, since AutoDock cannot compute an estimated energy for atoms outside the grid box. For a probe whose center of mass is near a grid box edge, some otherwise possible orientations would be disallowed because they would place one or more ligand atoms outside the grid, while for a probe nearer the center of the grid, all possible orientations would be permitted. Statistics taken from multiple docking runs indicated that target surface atoms within about 2Ǻ or less from a grid box edge are about five times less likely to be included in a docking site than target surface atoms that are 10Ǻ or more from the nearest grid box edge, with the probability increasing approximately linearly between those limits (and, curiously, dropping off somewhat above about 25Ǻ from the nearest grid box edge).

### Energy weighting of the probe docking poses

Given the hypothesis that peptide interactions with proteins are driven by relatively non-fastidious interactions between peptide moieties and protein surface moieties, it seems reasonable to suppose that interactions represented by probe docking poses with lower estimated ΔG (i.e. in which ΔG is a larger negative number) would have a higher probability of occurring than those corresponding to less energetic poses. For the results reported here, docking poses were weighted within each probe type (i.e. each distinct trimer sequence) by Boltzmann ratio according to the estimated docking energy reported by AutoDock for each pose, and then normalized over the various probe types so as to maintain the Boltzmann weighting within each probe type while weighting each probe type equally overall. This weighting strategy is not entirely satisfying, given that the collision rates and binding affinities of small probes may deviate substantially from Boltzmann weighting in terms of their contribution to the overall collision rate and binding affinity of a much larger peptide. Further, the genetic algorithm based search would be expected to return samples drawn disproportionately from the higher energy regions of the fitness space. Although somewhat relaxed values of the AutoDock search parameter settings were chosen relative to the number of rotatable bonds, so as to 'detune' the search and cause a wider diversity of solution poses to be sampled, it may be that some improvement in predictive performance could be obtained by reweighting on a distribution representing some compromise between a Boltzmann distribution and a uniform distribution. The equal weighting as between different probe types in maps based on multiple probe types is also conceptually questionable, but was found necessary because otherwise the highest affinity probes dominate the mapping to the point that any information about the likely position of less energetic moieties is completely lost.

### Computation of likelihood ratios

The VSPM software first computes the total number of expected “hits” available to be allocated over the set of target protein surface atoms. For each grid, the expected hit count is computed as $E_{g\,=}\,E_p\,P_g$, where $E_g$ is the expected total number of hits for all target surface atoms within grid g, $E_p$ is the average number of target surface atoms affected by each probe docking pose (as determined by statistics taken over all probe docking poses in all grids), and $P_g$ is the total number of probe docking poses within grid g.  This expected hit count is then allocated uniformly over all target surface atoms within the grid, after which the adjustment for edge effect is applied, decreasing the expected hit count for atoms ≤ 2Å from the nearest grid edge by a factor of 5.0, and for atoms between 2Å and 10Å by 10.0 divided by the distance from the nearest grid edge.  All counts are then normalized so that the total expected count for the grid $E_g$ remains unchanged. The individual atom counts over the multiple grids are summed, atom by atom, resulting in a count $E_a$ for each target surface atom representing the sum of that atom’s edge-effect-adjusted share of $E_g$ for each grid to which that atom belongs (keeping in mind that because the grids overlap, each atom may belong to more than one grid). These atom-by-atom counts $E_a$ are used as the denominator of the likelihood ratio for each atom.

A total expected hit count $E_s$ for all atoms of the target protein surface is computed as the sum of the individual surface atom expected hit counts $E_a$ over all surface atoms. A per-atom hit value $N_p$ is computed for each probe docking pose by allocating the total expected hit count $E_s$ over all probes in proportion to their respective Boltzmann weights (normalized to equalize the overall weight of each probe type). The allocated hit count for each probe docking pose (divided by the number of surface atoms affected by each probe) is then posted to the hit count $N_a$ for each target surface atom that is ≤ 4Å from any atom in the probe docking pose. Thus, the sum of $N_a$ over all target surface atoms is made to equal the sum of $E_a$ over all target surface atoms, and the likelihood ratio $M_a = N_a / E_a$ for any atom reflects that atom's propensity to be included in the docking sites of the probes. The set of target residues predicted as the binding site for the peptide can be taken as the set of residues containing at least one surface atom for which $M_a$ exceeds a suitable threshold, and the individual $M_a$ values can be color-mapped onto the surface for visualization.

### Evaluation of predictions

The model was evaluated based on its performance in predicting the set of residues belonging to the actual peptide binding site, for a set of eight protein-peptide complexes for which PDB structures of both the bound complex and the unbound form of the protein were available. The eight complexes were selected randomly from a dataset of 402 such complexes published by Petsalaki, et al. [235] A requirement was imposed that the selected structures must include a peptide of at least 8 and not more than 20 residues and a target protein of at least 90 residues, in which the peptide in the bound complex was substantially in contact with the target protein (many of the complexes in the Petsalaki dataset had only one or two residues in contact). Predictions were evaluated by comparing the set of residues predicted as belonging to the peptide binding site with the set of residues that actually contain at least one atom within 4Ǻ of at least one atom of the peptide, based on the positions per the PDB structure of the bound complex. For each of the eight proteins tested, predictions were made based on both the PDB structure of the bound protein-peptide complex with the peptide removed, and also based only on a PDB structure of the same protein in its unbound form. For each unbound structure, the set of residues predicted for the binding site was mapped to the identical residues of the PDB structure of the bound complex for comparison with the position of the bound peptide and extraction of statistics. Receiver operating characteristic (ROC) plots were prepared using Mathematica 7.0 [244].

### Benchmarking

Predictions were compared with similar predictions made using Pepsite [235], a peptide binding prediction algorithm based on statistics of spatial relationships drawn from PDB structures of folded proteins, and Optimal Docking Areas (ODA), an algorithm for locating protein binding hotspots based on hydrophobicity. Both algorithms are accessible via web servers [245, 246]. Pepsite returns five solutions each containing the predicted locations of the $C_\alpha$

atoms of several (but typically not all) of the residues of the bound peptide. To enable comparison with the VSPM predictions, Pepsite's prediction was taken as the set of target residues having any atom within 5Å of any of the atoms comprising any of the five solutions.  The Pepsite server does not accept peptide sequences longer than 10 residues; since five of the peptides in the test set are longer, for those complexes, first ten and last ten residues were submitted in separate runs, and the union of the two result sets was taken as Pepsite's prediction.  ODA returns a tabulation of values each representing the computed desolvation energy of a surface region, on a residue by residue basis, making it necessary to impose an arbitrary cutoff to identify a solution set. For each target protein, these values were ranked, and ODA's prediction was taken as that percentage of the lowest energy (i.e. highest negative energy) residues equalling (as nearly as possible) the percentage of surface residues encompassed by the VSPM prediction for the same structure.

### Applying geometric constraints

Finally, in an attempt to improve predictive accuracy, search was made for combinations of probe poses such that residues that are adjacent in the peptide sequence are placed at plausible spacings on the protein surface. Using probes comprising the set of overlapping trimer subsequences of the peptide of interest, the search algorithm attempted to identify solution sets each comprising exactly one probe docking pose of each probe type, chosen such that, within each set, adjacent overlapping trimer probes are positioned with their overlapping residues in reasonable register with each other and without steric interference between the non-overlapping portions of the chain. This amounts to a search of the space of all possible correctly sequence-ordered combinations of docking poses, with the geometric constraints being expressed, together with any other selection criteria, in the form of a fitness function. The potential complexity of the search is O( $(p \bullet g)^t$ ) where t is the number of overlapping trimers, p is the number of docking poses obtained per probe per grid, and g is the number of grids; thus, for a typical search over 25 probe poses in each of 10 grids for each of 10 overlapping trimer probes, the number of possible combinations, ignoring geometric constraints, is $250^{10}$ or approximately $10^{24}$. Three search algorithms were evaluated: dynamic programming, genetic algorithm, and tree search with pruning. Dynamic programming offers the advantages of high efficiency and guaranteed discovery of the global optimum; however, it attempts to build solution sequences recursively, and therefore must assume that the fitness contribution of recursively incorporated partial solutions is fixed and independent of the composition of the remainder of the solution. It was discovered that there is a tendency for docking poses, regardless of probe sequence, to prefer some of the same regions on protein surfaces, resulting in sterically impossible solution sets in which non-adjacent probe docking poses attempt to occupy the same space on the target surface. It was therefore necessary to take into account the overall steric plausibility of the solution set, which changes as each additional pose is added, making dynamic programming unsuitable. A genetic algorithm was tried, which has the advantage of placing no restrictions on the fitness function, but convergence was unacceptably slow and there is no guarantee of reaching the global optimum. The search space in question does have one important exploitable characteristic, however: if a requirement is imposed that each individual pair of adjacent overlapping probe docking poses in the solution set satisfy the given geometric proximity constraints, then it is possible to perform a breadth-first tree search in which the entire branch descending from any pair not satisfying such constraints can be pruned and eliminated.  Further, for the partial solution sets that remain, the overall steric plausibility can be evaluated and taken into account, and any partial branches failing to meet the steric criteria can also be pruned at each step. Thus, the set of partial solution sets satisfying the geometric and steric constraints remains at a reasonable size at each iteration, and the set of solution sets remaining after the last trimer in sequence has been added can then be evaluated and an optimal solution selected on the basis of any desired fitness criteria.  (The explanation for the poor convergence of the genetic algorithm likely lies in its inability to exploit the opportunity to permanently exclude any solution sets containing geometrically impermissible pairings, which, in the tree search algorithm, enormously reduces the size of the search space while preserving the guarantee of finding the global optimum.)

For the results here reported, a requirement was imposed that for a next-in-sequence docking pose to be added to a solution set, the following constraints must be satisfied:

(1) The sets S1 and S2 of target surface atoms within 5Å of any atom in the two overlapping residues shared by the probe pose to be added and the next preceding pose is determined. (Thus, for example, if the last pose of the growing solution set has the sequences ACD and the pose to be added has the sequence CDE, the set S1 is the set of target surface atoms "occupied" by residues C and D of the ACD pose and the set S2 is the set of target surface atoms "occupied" by residues C and D of the CDE pose.) A score in the range 0 to +100 is assigned based on the ratio of the accessible surface area shared by the two sets to the unshared accessible surface area, such that +100 signifies that S1 and S2 are identical and 0 signifies that S1 and S2 are entirely disjoint. A similar score is obtained representing the degree of surface overlap between the first residue of the trimer sequence to be added and the last residue of the penultimate trimer of the existing set. A weighted average is taken of these two scores.

(2) In a similar manner, a score is obtained representing the extent to which the surface atoms affected by the trimer pose to be added overlap the set of surface atoms affected by residues of the growing solution set that precede the residues of the pose to be added in the peptide sequence (i.e. the residues whose positions the added pose should not overlap). A high score here represents a solution set wherein the pose to be added is "doubling back" and attempting to occupy surface positions already occupied by a preceding part of the chain.

Because the method of assigning surface atoms as occupied by docking poses is imprecise, and it is possible for a surface atom to be within 5Å of atoms from two docking poses that are not in fact in collision, and because in performing the initial tree search it is preferable to err on the side of inclusion, prospective additions to a solution set are rejected if the pose to be added has more than 35% overlap with the non-corresponding residues of the preceding chain. An addition is also rejected if the weighted average overlap percentage between the first two residues of the pose to be added and the corresponding residues of the solution set is less than 35%.  As described below, other criteria are applied to select one or more particular solutions from the final (large) set of solutions that satisfy the threshold geometric criteria; the overlap cutoffs affect only the liberality of the pruning, not the selection of the ultimate solution.

Each iteration of the search corresponds to the addition of one trimer position from the peptide of interest.  Thus, for example, for a peptide ACDEFGHI, for which sample docking poses have been generated for the trimers ACD, CDE, DEF, EFG, FGH, and GHI, the first iteration evaluates all possible combinations of ACD poses and CDE poses, and retains as tentative solution sets any pairs of ACD + CDE poses that satisfy the foregoing constraints. In the next iteration, each pairing of each of the retained solutions with each DEF pose is similarly evaluated, and each ACD + CDE + DEF set satisfying the constraints is retained. The search proceeds in this manner until the end of the sequence is reached or until no additions satisfying the constraints can be made. To account for the possibility that only part of the peptide chain may bind, the search is repeated starting with the second and each subsequent trimer in sequence, and any partial solutions longer than a specified minimum are retained. The end result is a large set of tentative solutions, each solution comprising a full or partial ordered series of unique docking poses that together correspond to the peptide sequence and satisfy the given constraints. Because the search algorithm prunes any branches that do not satisfy the constraints, and retains and attempts to extend all partial solutions that do, the final set of solutions is guaranteed to contain all possible combinations that satisfy the constraints (subject to tightening of the constraints in the event that the set of solutions becomes too large). The intent is that all solutions are retained in this set except those that are obviously impossible, as where probes representing adjacent parts of the peptide sequence are so far apart that they could not plausibly belong to a contiguous chain.

The solutions in the set of tentative solutions are then scored and ranked according to a linear combination of two equally weighted factors: (1) the aggregate predicted energy of the probes comprising the solution, and (2) the extent to which the solution does or does not lie in the regions of the target surface most likely to belong to the binding site, as measured by the target surface atom likelihood ratios already described.  Each score is scaled by the length of the predicted bound chain segment.

## Results and discussion

In the following sections the results of VSPM-based binding site predictions are presented for a test set of eight peptide-protein complexes for which PDB structures are available for both the bound complex and the unbound protein (see Table 8), first using relatively generic probes and then using probes specifically derived from the peptide sequences of interest to improve predictive accuracy, and address factors affecting the performance of each of the strategies described. These results are compared to those obtained from two other published approaches for identifying likely interaction sites, and it is shown that the VSPM method correctly finds peptide binding sites in several cases where the other methods fail. An evaluation is then made of the potential for improving the solution by incorporating positional constraints derived from the known peptide sequence. Finally, the discussion turns to issues affecting VSPM-based predictions generally, and to the implications of the reported results as they relate to the mechanism of peptide binding.

### Maps derived from simple, single-amino acid probes are informative regarding spatial distribution of binding preferences.

As expected, the use of probes having different chemical characteristics results in maps having quite different patterns of predicted likelihood of interaction. Figure 30(a)-(d) compares the maps obtained using a positively charged probe (an Ala-Lys-Ala tripeptide), a negatively charged probe (Ala-Glu-Ala), a hydrophobic probe (Ala-Ile-Ala), and an aromatic probe (Ala-Trp-Ala), on opposite faces of a representative protein structure.

| PDB ID | Peptide Sequence | Protein Chain | Chain Length | Description |
|---|---|---|---|---|
| 1MK9 | HMWDTANNPLYKEA | F, H | 378 | Integrin beta-3 talin chimera |
| 1TWB | ACNDENYA | B | 106 | Hydrolase |
| 1U8T | SILSQAEIDALLN | B | 122 | Signalling protein |
| 1VWG | CHPQGPPC | B | 121 | Streptavidin |
| 1X8S | HREMAVDCP | A | 98 | PAR-6 PDZ domain |
| 2BP3 | TFRSSLFLWVR | B | 91 | Filamin actin binding protein |
| 2FF6 | ETNEKNPLPDK | A | 360 | Structural protein |
| 2IVZ | GASDGSGWSSENNPWG | D | 387 | TOLB |

Table 8. Description of the eight peptide-protein complexes comprising the test set.

**Protein surface interaction maps obtained using simple, single-amino acid probes that correspond to the amino acids present in a particular peptide are together predictive of the residues comprising the binding site of the peptide.**

When maps of this kind are compared to the actual binding loci of the corresponding residues in peptides in bound complexes with proteins, the correspondence is considerably better than would be expected by chance. For the 8 peptide-protein complexes evaluated, the protein surface residues having a likelihood ratio > 3.0 on interrogation with a probe in the form Ala-X-Ala comprised 40.4 percent of the protein surface residues actually in contact with residue(s) X of the bound peptide. Maps of this kind would be expected to over-predict, giving a relatively high false positive rate (here, 14.1 percent overall), because each PDB structure contains only a single bound peptide in a single locus, but there may often be multiple surface loci that show interactivity with a given probe type (see Figure 30). The false positive rate can of course be reduced by imposing a more stringent likelihood ratio cutoff, but doing so reduces sensitivity.  Figure 31 (dashed line) shows a receiver operating characteristic (ROC) plot of the relationship between true positive rate (TPR) and false positive rate (FPR) for the 63 Ala-X-Ala-based predictions of individual peptide residue positions in the 8 complexes.

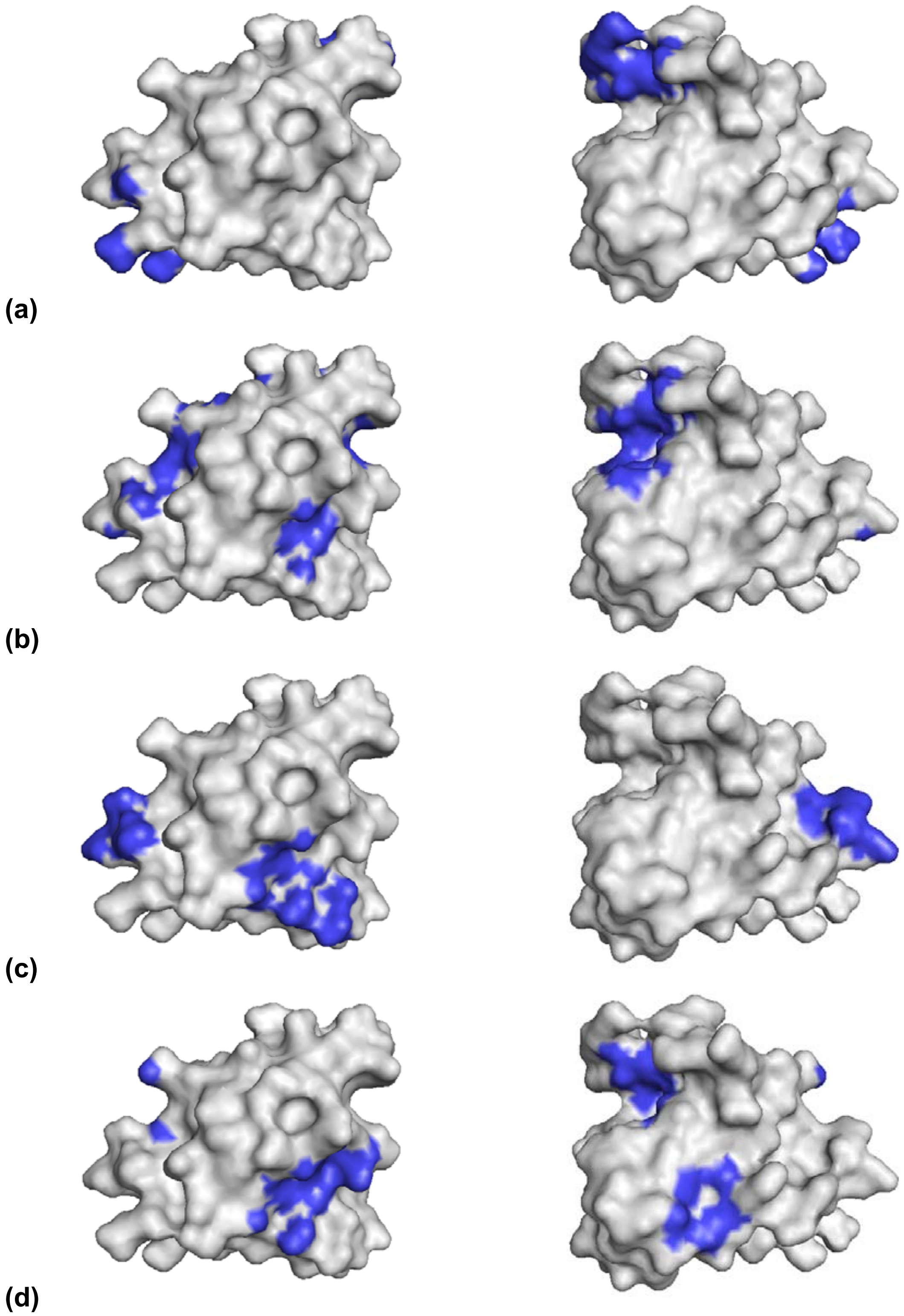
(a)
(b)
(c)
(d)

Figure 30 (preceding page). VSPM map of PDB:1TWB, generic probes. Opposite (180 degree rotated) faces of PDB:1TWB. Dark regions have  likelihood ratio > 3.0for interaction with probes (a) Ala-Asp-Ala, (b) Ala-Ile-Ala, (c) Ala-Lys-Ala, and (d) Ala-Trp-Ala.

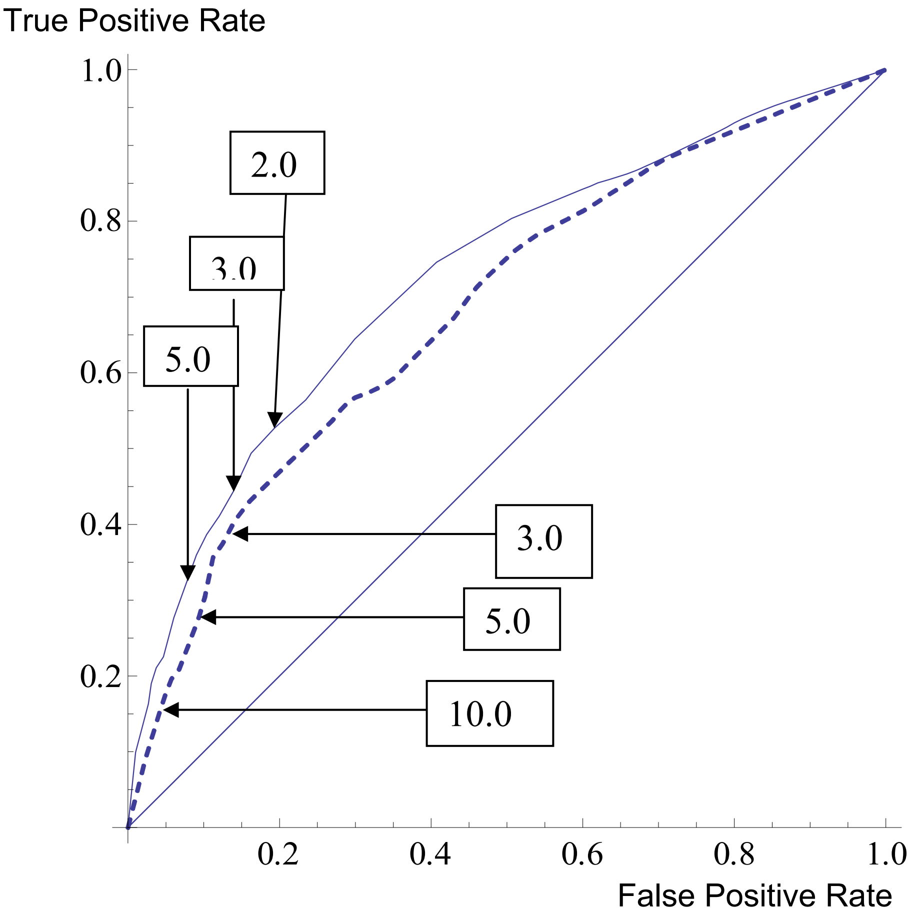

Figure 31. Receiver operating characteristic curves, Ala-X-Ala probes vs. trimer subsequence probes.  Receiver operator characteristic curve for predictions of residues interacting with Ala-Res-Ala probes representing residues present in bound peptides (dashed line) or trimer probes representing all trimer subsequences present in bound peptide (solid line), for 8 protein-peptide complexes, reflecting TP rate vs. FP rate at various likelihood ratio cutoffs as shown.

When interaction maps are computed based on combined interrogation of the surface using multiple single-amino acid probes, they typically highlight one or more larger regions that might be expected to have a relatively higher affinity for binding by a peptide that contains residues corresponding to those of the probes. Figure 32 compares, for the eight protein-peptide complexes of the test set, the known position of the peptide from the bound complex with the surface regions having a likelihood ratio > 3.0 for interaction with simple Ala-X-Ala probes with X representing each of the residues of the peptide. In seven of the eight cases, the predicted loci correspond to at least part of the binding site of the bound peptide. Again as expected, additional loci away from the actual binding site are also predicted, some of which are in parts of the molecules not visible in Figure 32.

**Binding site prediction accuracy is improved using probes that take into account flanking residues in the peptide of interest.**

It seems reasonable to hypothesize that predictive accuracy should be improved to whatever extent additional information about the specific ligand of interest can be incorporated into the design of the probes. In particular, probes that more accurately model the microenvironment resulting from the interactions between peptide residues and their neighboring residues should improve results. For sets of probes in the form A-X-B, consisting of all of the possible contiguous trimer subsequences from the bound peptide, the protein surface residues having a likelihood ratio > 3.0 on interrogation with the probe comprised, in aggregate, 44.4 percent of the protein surface residues actually in contact with the corresponding residue X of the bound peptide, with a false positive rate of 13.9. The ROC relationship is shown in Figure 31 (solid line).

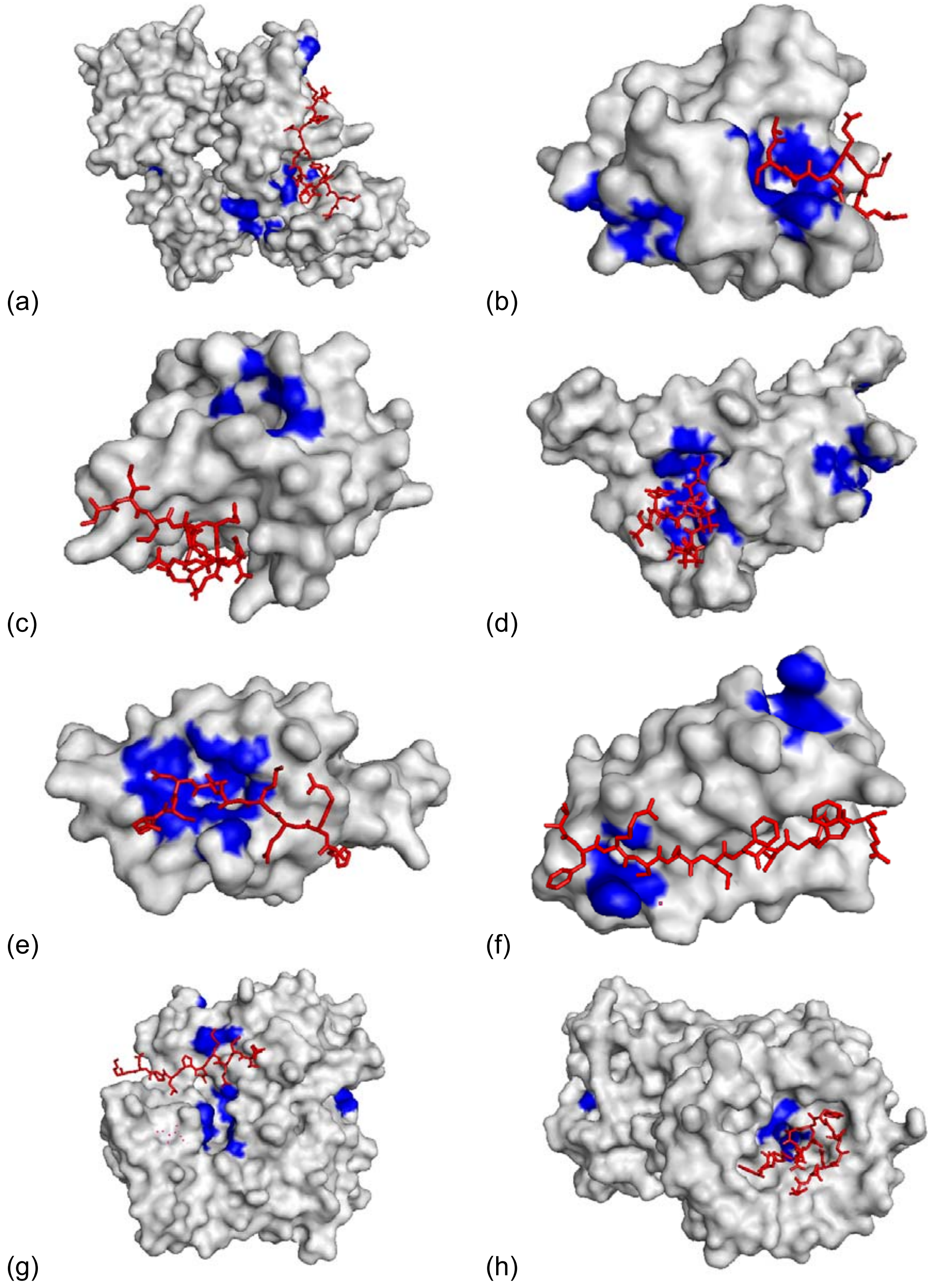
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

Figure 32 (Preceding page). VSPM maps of test set, Ala-X-Ala probes. Combined surface interaction maps based on probes Ala-residue-Ala. (a) PDB:1MK9, (b) PDB:1TWB, (c) PDB:1U8T, (d) PDB:1VWG, (e) PDB:1X8S, (f) PDB:2BP3, (g) PDB:2FF6, (h) PDB:2IVZ. Blue: residues with interaction likelihood ratio > 3.0. Red: peptide from bound complex.

**The combined surface maps using peptide trimer probes extracted from the sequence of the bound peptide are predictive of the binding loci of the peptides in the bound complexes.**

When the surface interrogation datasets from probes consisting of all of the possible contiguous trimer subsequences from the bound peptide are combined to produce a single protein surface map, and again setting the likelihood ratio operating point at 3.0 (i.e. considering the predicted binding site to comprise all surface residues containing at least one atom whose likelihood ratio is > 3.0), at least 50 percent of the actual binding site residues are predicted for four of the eight complexes, with a false positive rate less than 10 percent (Figure 33 and Figure 34, black lines; black arrows indicate operating points). At least 25 percent of the true binding site residues are predicted for seven out of eight of the complexes, again with a false positive rate less than 10 percent.

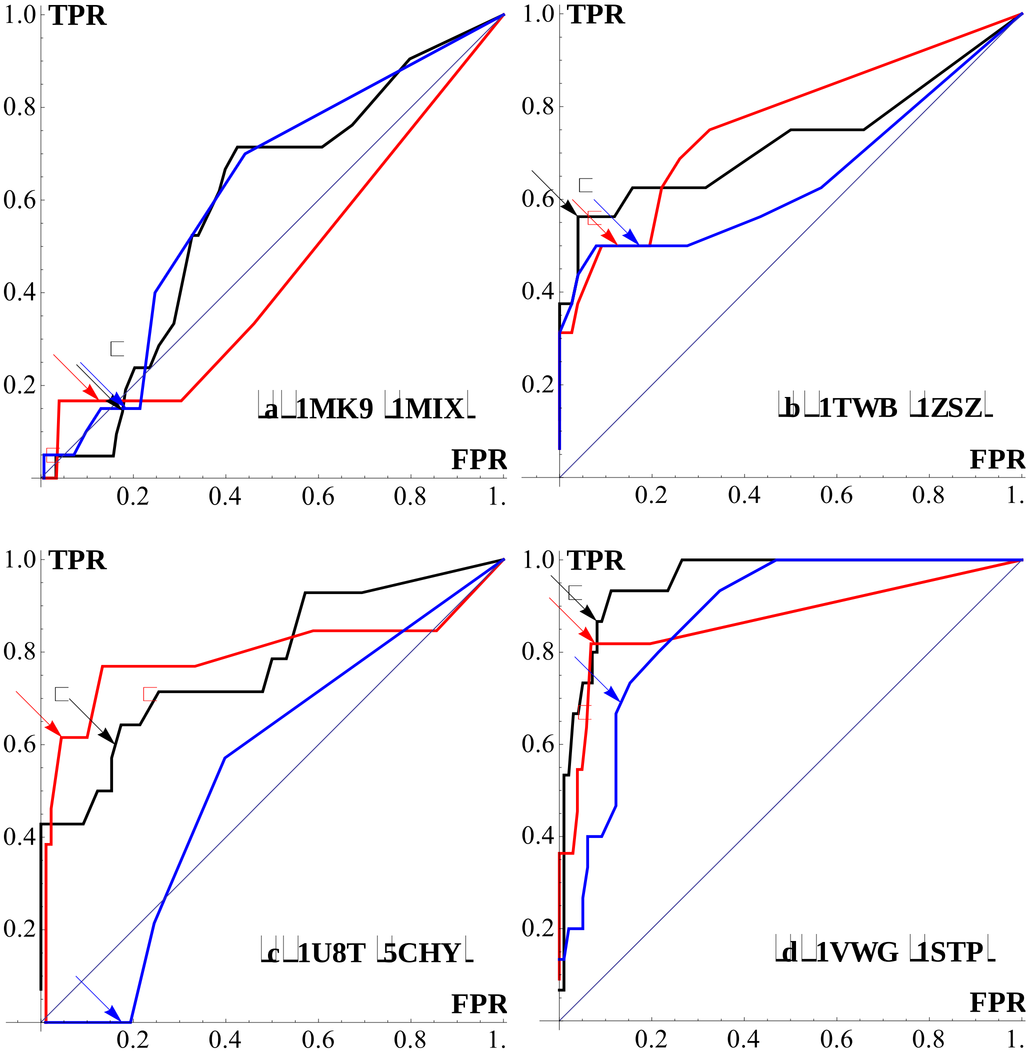

Figure 33. Receiver operating characteristic curves, contiguous trimer probes. Receiver operating characteristic curves for (a) PDB:1MK9, (b) PDB:1TWB, (c) PDB:1U8T, and (d) PDB:1VWG, each comparing true positive rate (vertical axis) vs. false positive rate (horizontal axis) for the set of residues predicted as comprising the peptide binding site, as function of likelihood ratio threshold chosen. Predictions based on contiguous trimer probes using bound structure of protein (black lines), unbound form of same protein (red lines, PDB ID shown in parenthesis), and Ala-X-Ala probes applied to bound form (blue lines). Arrows: operating point based on likelihood ratio threshold of 3.0. ⎛ / ⎛ : TPR vs FPR values for highest ranking solution with geometric constraints shown for bound and unbound structure, respectively.

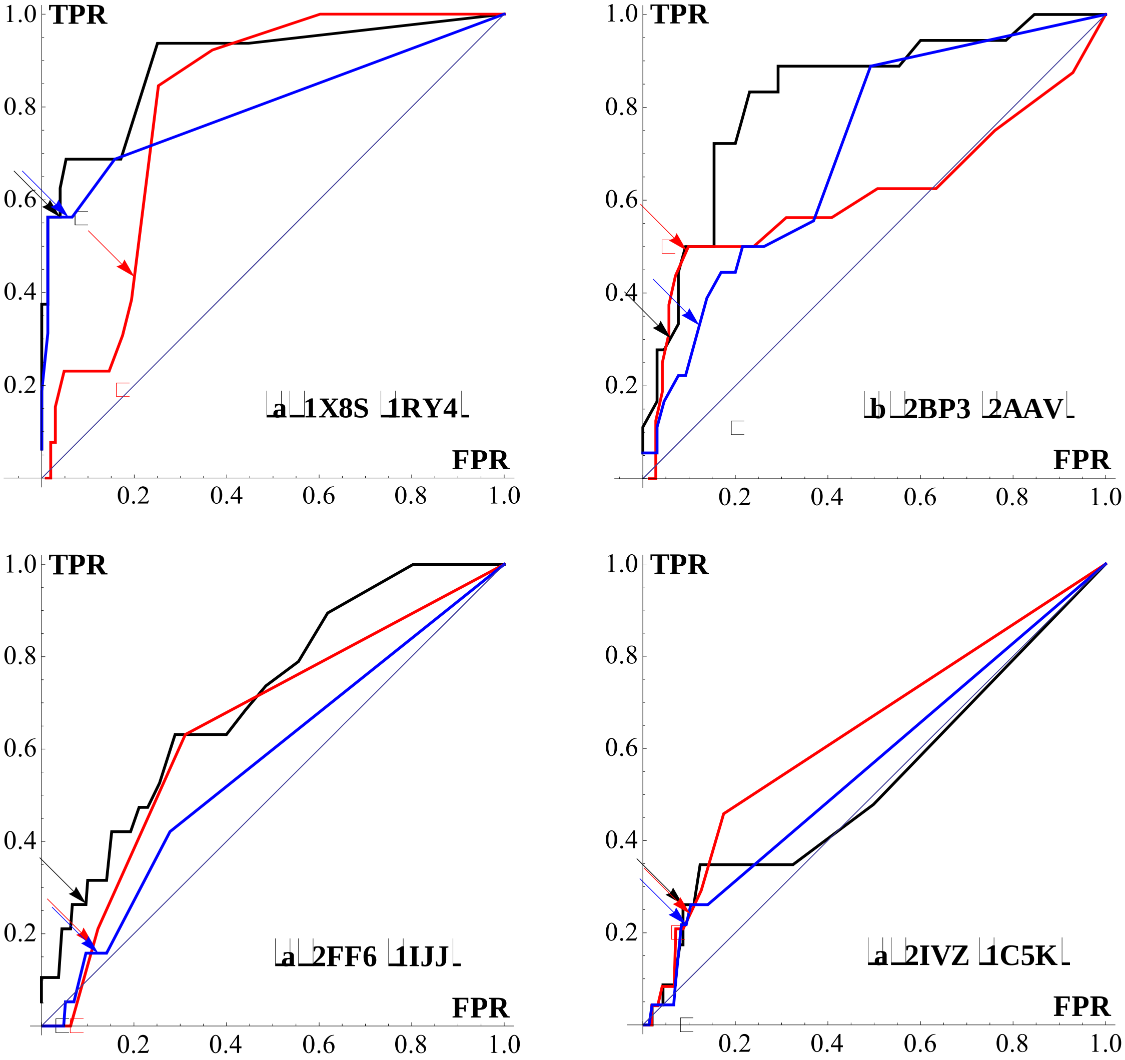

Figure 34. Receiver operating characteristic curves, contiguous trimer probes. ROC curves for (a) PDB:1X8S, (b) PDB:2BP3, (c) PDB:2FF6, (d) PDB:2IVZ, each comparing true positive rate (vertical axis) vs. false positive rate (horizontal axis) for the set of residues predicted as comprising the peptide binding site, as function of likelihood ratio threshold chosen. Predictions based on contiguous trimer probes using bound structure of protein (black lines), unbound form of same protein (red lines), and Ala-X-Ala probes applied to bound form (blue lines). Arrows: operating point based on likelihood ratio threshold of 3.0. □ / □ : TPR vs FPR values for highest ranking solution with geometric constraints shown for bound and unbound structure, respectively.

As shown in Table 9, for these predictions, the overall mean positive predictive value (PPV) is 0.44, TPR is 0.45, and specificity is 0.91. As is apparent comparing these ROC plots with the corresponding ROC plots for similar predictions based on Ala-X-Ala probes (Figure 33 and Figure 34, blue lines, blue arrows indicate operating points; see Figure 31, solid line, for composite plot), use of the more specifically tailored probes considerably increased predictive performance (PPV of 0.44 as compared to 0.28 for the Ala-X-Ala probes at likelihood ratio cutoff of 3.0; see Table 2). Figure 35 and Figure 36 show molecular images of each of the eight target proteins reflecting the predicted mapping in comparison to the known position of the bound peptide, colored to reflect the regions corresponding to true positive (red**),** false positive (yellow), true negative (blue), and false negative (green) predictions. See the supplemental files (Appendix 1) for Pymol session files corresponding to each mapping.

| **Operating Point (Likelihood Ratio Cutoff)** | | **2.5** | | | | **3** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **TPR** | **FPR** | **PPV** | **SPC** | **TPR** | **FPR** | **PPV** | **SPC** |
| **1MK9** | Bound | 0.24 | 0.23 | 0.12 | 0.77 | 0.14 | 0.18 | 0.1 | 0.82 |
| (1MIX) | Unbound | 0.17 | 0.15 | 0.04 | 0.85 | 0.17 | 0.13 | 0.04 | 0.87 |
| | AXA | 0.16 | 0.22 | 0.09 | 0.78 | 0.15 | 0.18 | 0.1 | 0.82 |
| **1TWB** | Bound | 0.56 | 0.08 | 0.61 | 0.92 | 0.56 | 0.04 | 0.75 | 0.96 |
| (1ZSZ) | Unbound | 0.5 | 0.17 | 0.38 | 0.83 | 0.5 | 0.13 | 0.45 | 0.87 |
| | AXA | 0.5 | 0.23 | 0.32 | 0.77 | 0.5 | 0.17 | 0.39 | 0.83 |
| **1U8T** | Bound | 0.64 | 0.19 | 0.33 | 0.81 | 0.6 | 0.16 | 0.35 | 0.84 |
| (5CHY) | Unbound | 0.62 | 0.06 | 0.59 | 0.94 | 0.62 | 0.05 | 0.66 | 0.95 |
| | AXA | 0.04 | 0.2 | 0.02 | 0.8 | 0 | 0.18 | 0 | 0.82 |
| **1VWG** | Bound | 0.87 | 0.09 | 0.59 | 0.91 | 0.87 | 0.08 | 0.62 | 0.92 |
| (1STP) | Unbound | 0.82 | 0.08 | 0.52 | 0.92 | 0.82 | 0.08 | 0.53 | 0.92 |
| | AXA | 0.74 | 0.16 | 0.42 | 0.84 | 0.69 | 0.13 | 0.44 | 0.87 |
| **1X8S** | Bound | 0.63 | 0.04 | 0.77 | 0.96 | 0.56 | 0.04 | 0.75 | 0.96 |
| (1RY4) | Unbound | 0.57 | 0.22 | 0.25 | 0.78 | 0.43 | 0.2 | 0.21 | 0.8 |
| | AXA | 0.58 | 0.08 | 0.62 | 0.92 | 0.56 | 0.06 | 0.68 | 0.94 |
| **2BP3** | Bound | 0.5 | 0.12 | 0.53 | 0.88 | 0.3 | 0.06 | 0.59 | 0.94 |
| (2AAV) | Unbound | 0.5 | 0.14 | 0.44 | 0.86 | 0.49 | 0.09 | 0.54 | 0.91 |
| | AXA | 0.44 | 0.17 | 0.42 | 0.83 | 0.33 | 0.12 | 0.42 | 0.88 |
| **2FF6** | Bound | 0.32 | 0.14 | 0.14 | 0.86 | 0.26 | 0.1 | 0.16 | 0.9 |
| (1IJJ) | Unbound | 0.29 | 0.16 | 0.11 | 0.84 | 0.18 | 0.11 | 0.09 | 0.89 |
| | AXA | 0.16 | 0.14 | 0.07 | 0.86 | 0.16 | 0.12 | 0.09 | 0.88 |
| **2IVZ** | Bound | 0.26 | 0.09 | 0.18 | 0.91 | 0.26 | 0.09 | 0.19 | 0.91 |
| (1C5K) | Unbound | 0.26 | 0.11 | 0.15 | 0.89 | 0.24 | 0.1 | 0.15 | 0.9 |
| | AXA | 0.24 | 0.1 | 0.16 | 0.9 | 0.22 | 0.09 | 0.15 | 0.91 |
| **Mean** | Bound | 0.5 | 0.12 | 0.41 | 0.88 | **0.45** | **0.09** | **0.44** | **0.91** |
| | Unbound | 0.47 | 0.14 | 0.31 | 0.86 | **0.43** | **0.11** | **0.33** | **0.89** |
| | AXA | 0.36 | 0.16 | 0.26 | 0.84 | **0.33** | **0.13** | **0.28** | **0.87** |

Table 9. Statistics for VSPM peptide binding site predictions using peptide trimer probes. True positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), and specificity (SPC) for predictions of surface residues belonging to peptide binding sites in the peptide-protein complexes from the PDB ID's indicated (bound complexes in bold, unbound forms of protein target in parentheses), at likelihood ratio cutoffs 2.5 and 3.0, based on composite protein surface map from the set of contiguous trimers derived from the sequence of each peptide.

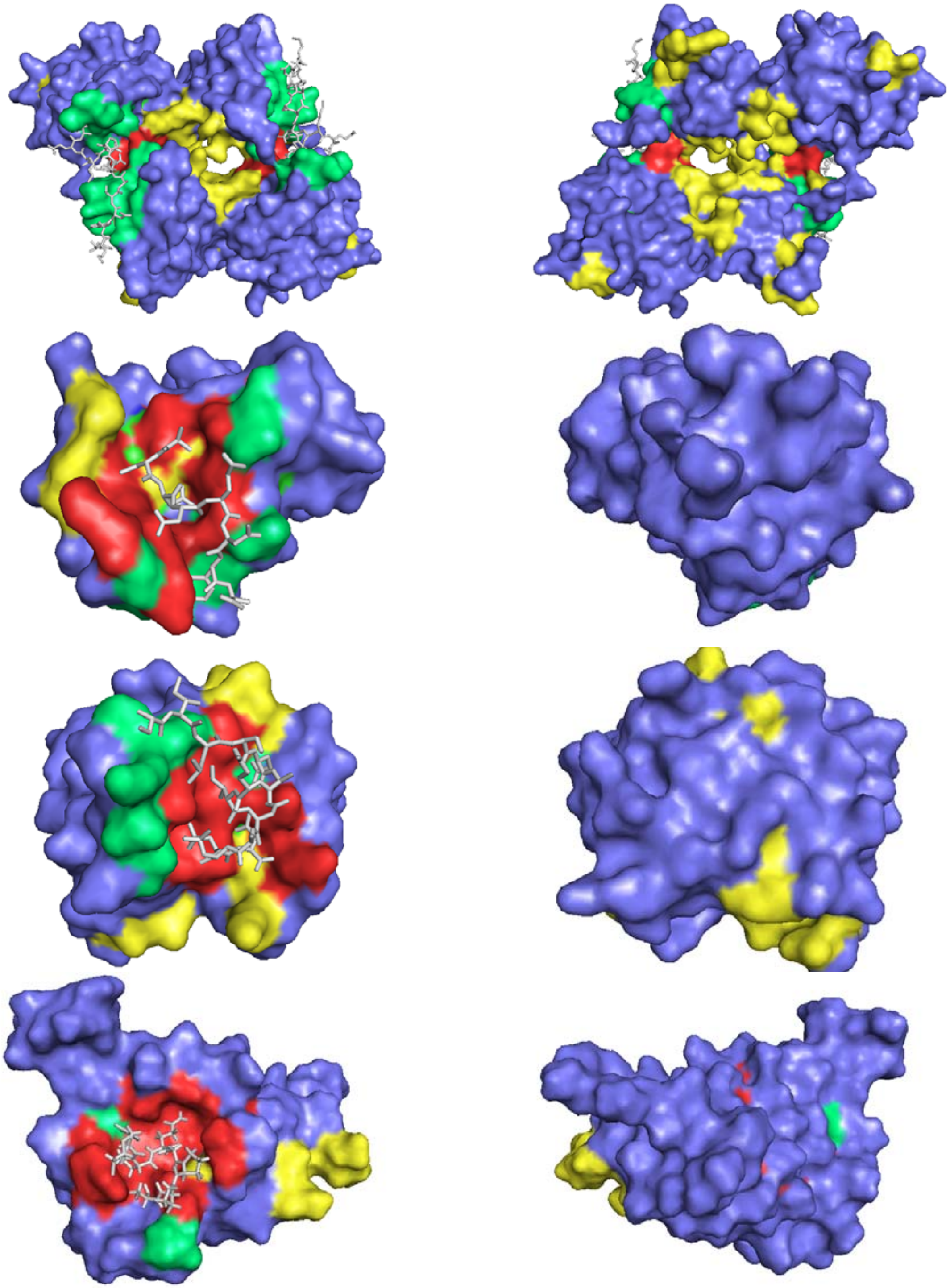

Figure 35 (Preceding page). VSPM maps of test set, contiguous trimer probes**.** Surface map images for peptide binding loci predictions based on probes comprising all trimer sequences present in peptides for the PDB peptide-protein complexes shown. Coloring is by residue, red: true positive; yellow: false positive; blue: true negative; green: false negative. (a) PDB:1MK9 (b) PDB:1TWB (c) PDB:1U8T (d) PDB:1VWG.

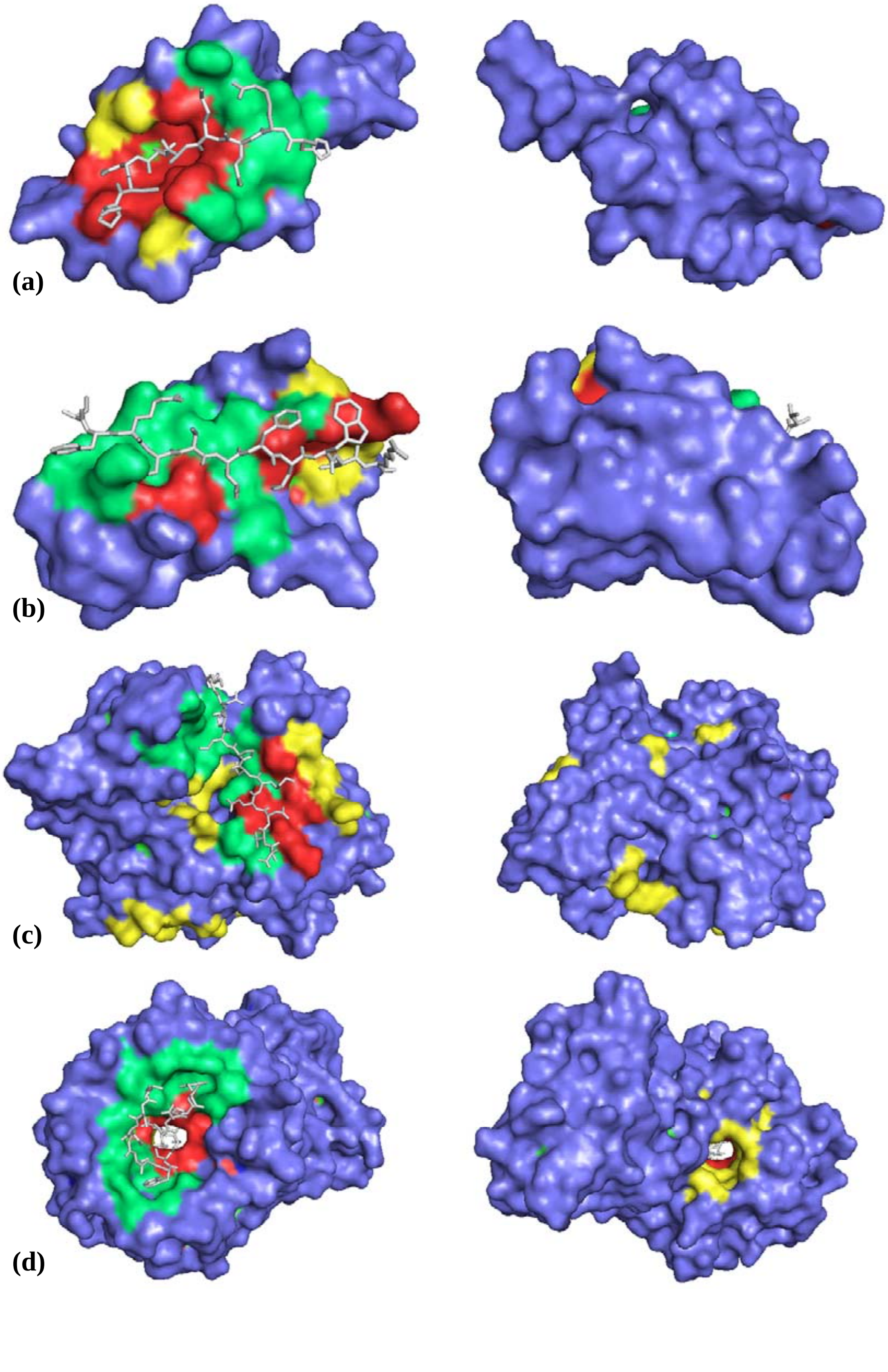
(a)
(b)
(c)
(d)

Figure 36 (Preceding page). VSPM maps of test set, contiguous trimer probes. Surface map images for peptide binding loci predictions based on probes comprising all trimer sequences present in peptides for the PDB peptide-protein complexes shown. Coloring is by residue, red: true positive; yellow: false positive; blue: true negative; green: false negative. (a) PDB:1X8S (b) PDB:2BP3 (c) PDB:2FF6 (d) PDB:2IVZ.

### Comparison with Pepsite and ODA methods

The foregoing results from the VSPM method were compared with data obtained using two other methods that can be adapted to make predictions of this kind and that appear to be representative of the current state of the art: the "Pepsite" algorithm of Petsalaki et al., and the "Optimal Docking Area" analysis of Fernandez-Recio et al. For the eight complexes in the test set, the VSPM method performed well by comparison, achieving an overall mean PPV of 0.44 as compared to 0.16 for ODA and 0.12 for Pepsite, with a mean TPR of 0.45 and FPR of 0.09, as compared to 0.25 and 0.16 for ODA and 0.21 and 0.20 for Pepsite. Statistics for each of the eight complexes are shown in Table 10, and molecular images showing the predicted binding loci are shown in Figure 37 and Figure 38 (ODA) and Figure 39 and Figure 40 (Pepsite). ODA and Pepsite may be disadvantaged in this comparison in that neither program is designed specifically for the task to which they were put, and obtaining the predictions shown required certain adaptations and parameter choices (described in the methods section) that the creators of those programs might well have made differently. Both programs nevertheless performed admirably, and the comparison is offered merely by way of providing an objective benchmark for assessing the VSPM method. Indeed, although the VSPM method succeeded in identifying binding sites in three of the complexes where ODA and Pepsite failed to do so (TPR of 56%, 30%, and 26%, respectively for complexes PDB:1TWB, PDB:2BP3, and PDB:2FF6, compared to  0%, 0%, and 5%, respectively, for ODA and 0%, 0%, and 0% for Pepsite), ODA and Pepsite located the binding site in one complex where the VSPM method performed poorly (TPR of 62% for ODA and 38% for Pepsite vs. 14% for the VSPM method on PDB:1MK9), and ODA outperformed the VSPM method on one other complex (39% vs. 26% on PDB:2IVZ).

|  |  | TPR | FPR | PPV | SPC |
|---|---|---|---|---|---|
| **1MK9** | VSPM | 0.14 | 0.18 | 0.10 | 0.82 |
|  | ODA | 0.62 | 0.13 | 0.39 | 0.88 |
|  | Pepsite | 0.38 | 0.20 | 0.21 | 0.84 |
| **1TWB** | VSPM | 0.56 | 0.04 | 0.75 | 0.96 |
|  | ODA | 0.00 | 0.17 | 0.00 | 0.84 |
|  | Pepsite | 0.00 | 0.09 | 0.00 | 0.92 |
| **1U8T** | VSPM | 0.60 | 0.16 | 0.35 | 0.84 |
|  | ODA | 0.21 | 0.24 | 0.11 | 0.78 |
|  | Pepsite | 0.21 | 0.29 | 0.10 | 0.71 |
| **1VWG** | VSPM | 0.87 | 0.08 | 0.62 | 0.92 |
|  | ODA | 0.33 | 0.17 | 0.23 | 0.84 |
|  | Pepsite | 0.53 | 0.31 | 0.21 | 0.78 |
| **1X8S** | VSPM | 0.56 | 0.04 | 0.75 | 0.96 |
|  | ODA | 0.38 | 0.18 | 0.30 | 0.93 |
|  | Pepsite | 0.44 | 0.16 | 0.37 | 0.88 |
| **2BP3** | VSPM | 0.30 | 0.06 | 0.59 | 0.94 |
|  | ODA | 0.00 | 0.15 | 0.00 | 0.85 |
|  | Pepsite | 0.00 | 0.25 | 0.00 | 0.78 |
| **2FF6** | VSPM | 0.26 | 0.10 | 0.16 | 0.90 |
|  | ODA | 0.05 | 0.14 | 0.03 | 0.86 |
|  | Pepsite | 0.00 | 0.13 | 0.00 | 0.91 |
| **2IVZ** | VSPM | 0.26 | 0.09 | 0.19 | 0.91 |
|  | ODA | 0.39 | 0.13 | 0.19 | 0.88 |
|  | Pepsite | 0.09 | 0.15 | 0.04 | 0.88 |
| **Mean** | VSPM | **0.45** | **0.09** | **0.44** | **0.91** |
|  | ODA | **0.25** | **0.16** | **0.16** | **0.86** |
|  | Pepsite | **0.21** | **0.20** | **0.12** | **0.84** |

Table 10. Comparison of VSPM predictions with Pepsite and ODA methods. Comparison of predictions for VSPM, the Optimal Docking Area method of Fernandez-Recio et al. (ODA) [13], and the Pepsite method of Petsalaki et al. [12, 26]

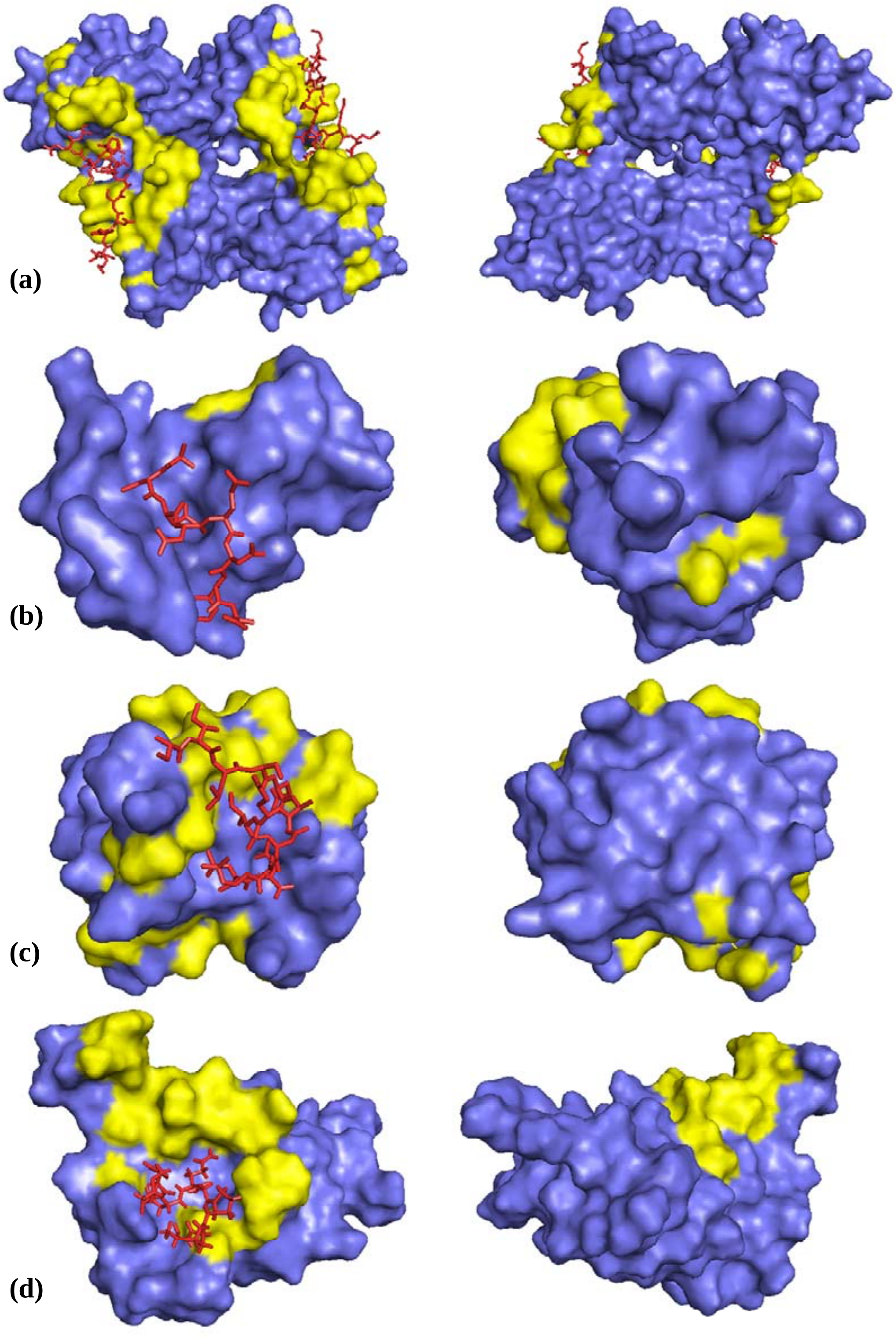
(a)
(b)
(c)
(d)

Figure 37 (Preceding page). Mapping of predictions of ODA method applied to test set. Predictions by Optimal Docking Areas method [13] of binding 'hot spots' (yellow) for the proteins shown. Actual bound position of the each peptide is shown in red. (a) PDB:1MK9 (b) PDB:1TWB (c) PDB:1U8T (d) PDB:1VWG.

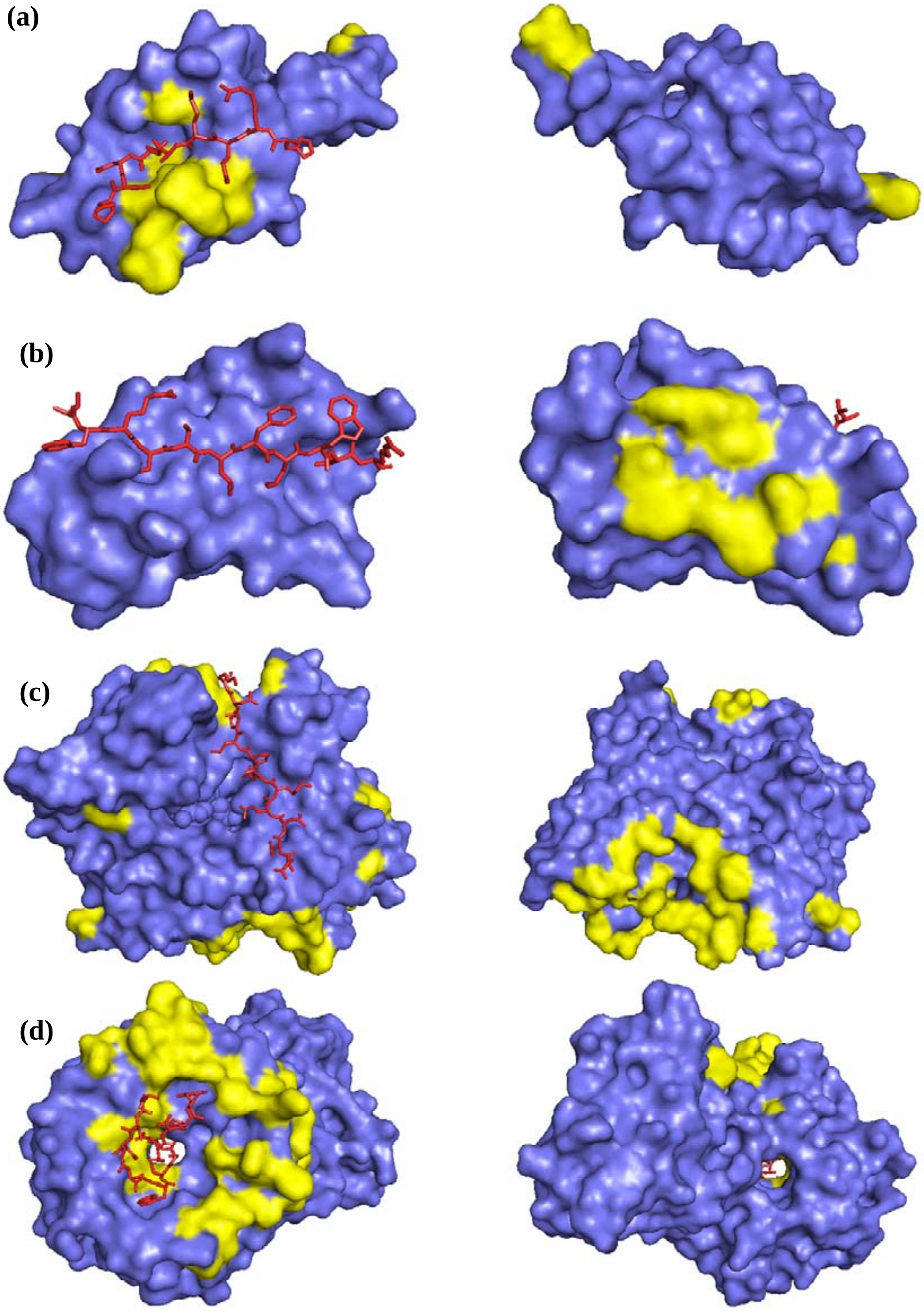
(a)
(b)
(c)
(d)

Figure 38 (Preceding page). Mapping of predictions of ODA method applied to test set. Predictions by Optimal Docking Areas method [13] of binding ‘hot spots’ (yellow) for the proteins shown.  Actual bound position of the each peptide is shown in red. (a) PDB:1X8S (b) PDB:2BP3 (c) PDB:2FF6 (d) PDB:2IVZ.

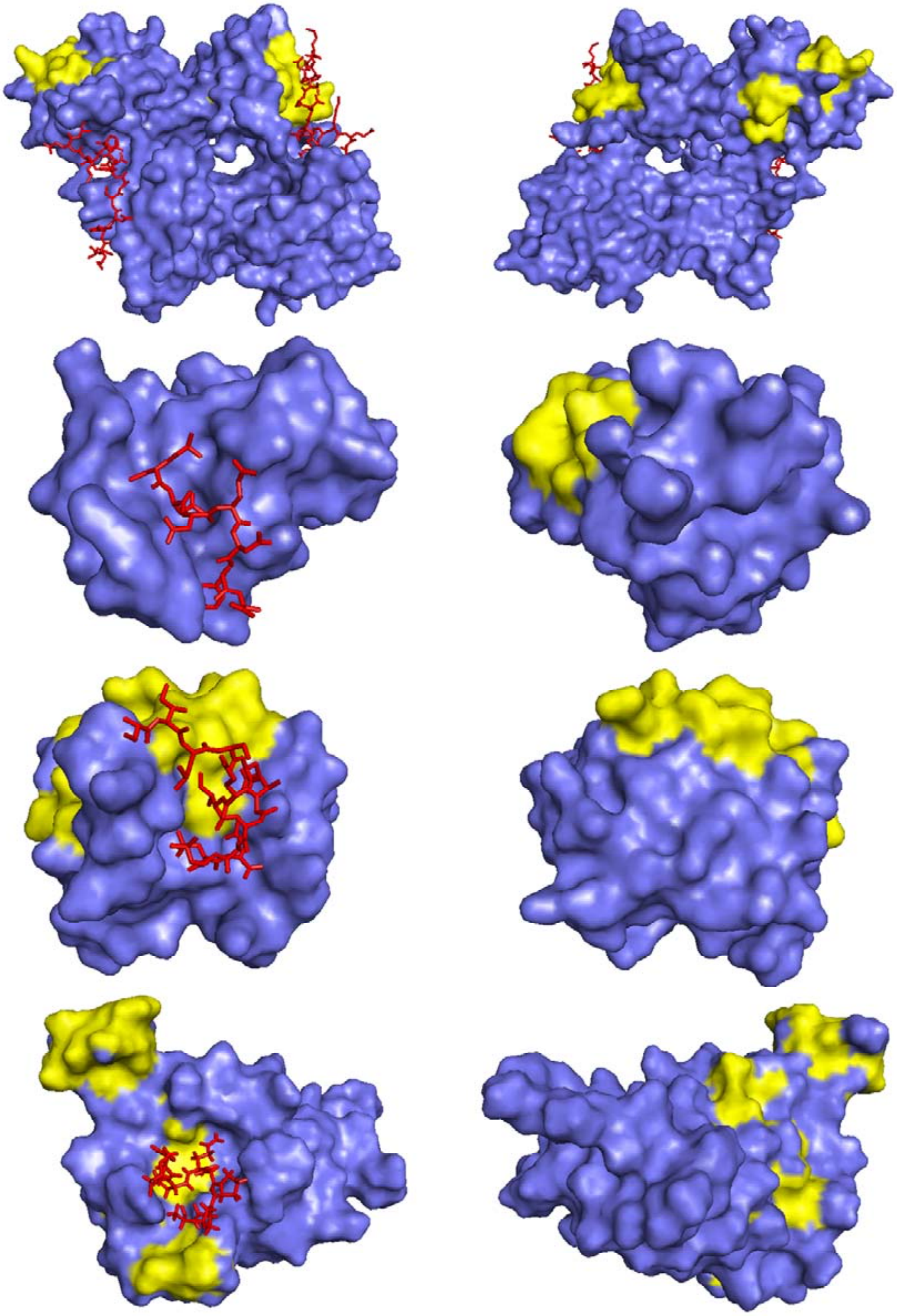

Figure 39 (Preceding page). Mapping of predictions of Pepsite method applied to test set. Predictions (yellow) by Pepsite algorithm [12, 26] of binding sites of peptides in the complexes shown. Actual bound positions are shown in red. (a) PDB:1MK9 (b) PDB:1TWB (c) PDB:1U8T (d) PDB:1VWG.

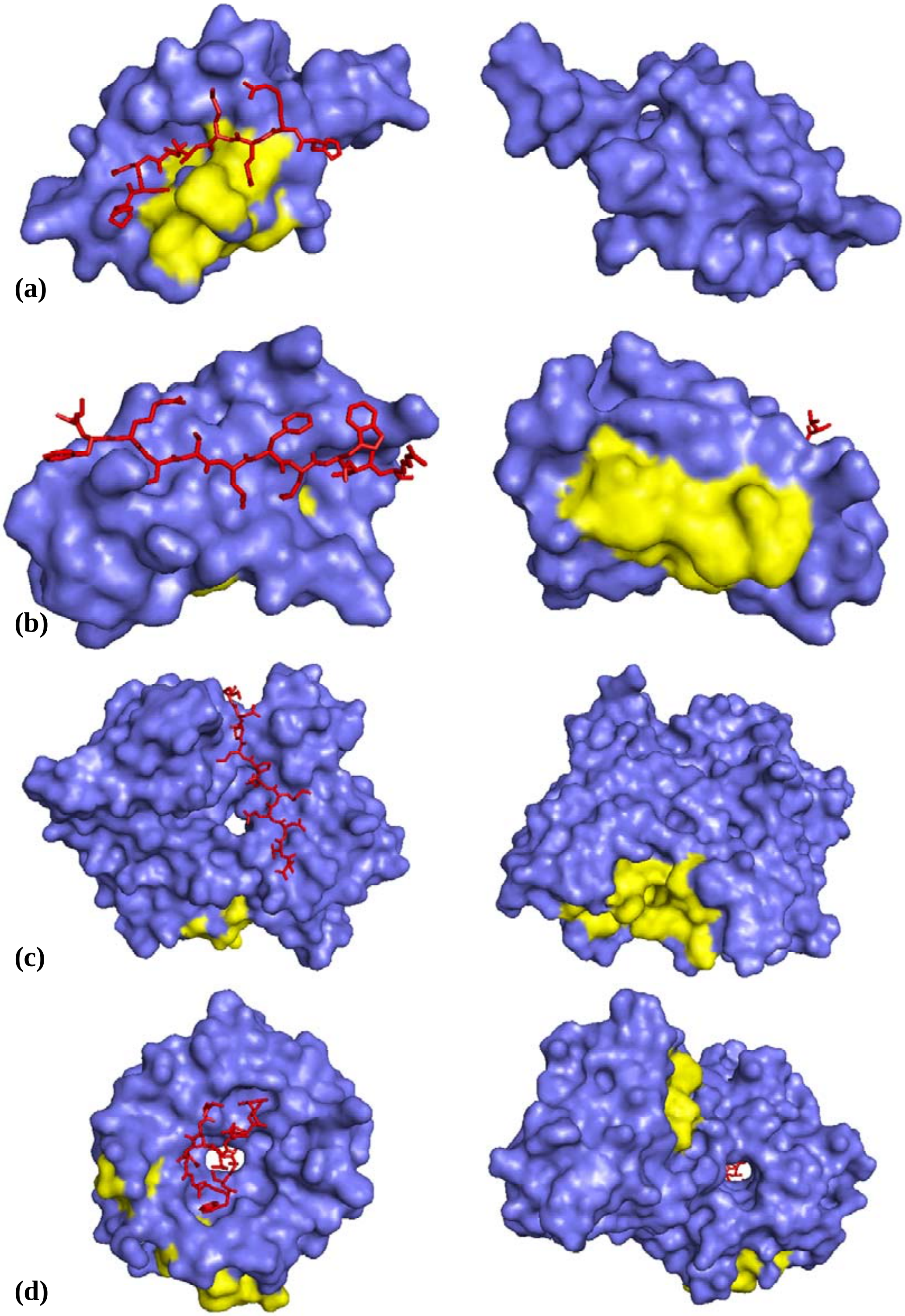
(a)
(b)
(c)
(d)

Figure 40 (Preceding page). Mapping of predictions of Pepsite method applied to test set. Predictions (yellow) by Pepsite algorithm [12, 26] of binding sites of peptides in the complexes shown. Actual bound positions are shown in red. (a) PDB:1X8S (b) PDB:2BP3 (c) PDB:2FF6 (d) PDB:2IVZ.

**Predictions are robust with respect to differences between bound and unbound structures.**

Absent a solved structure of the peptide-protein complex of interest – which, if available, would make binding site prediction unnecessary – any “real life” predictions would have to be based on unbound structures of the target, and these can be expected to differ from the bound form. Protein surfaces often change upon binding by a peptide, particularly in the binding region; even the highest resolution structural models are to some degree inaccurate; and proteins often contain relatively unstructured regions.  The predictions described to this point have all been made on the basis of analyzing the target protein in its exact bound conformation. To assess the sensitivity of the VSPM method to error in the protein structure being mapped, maps were constructed based on the unbound forms of the eight target proteins comprising the test set. These maps were computed using as probes, for each target, the set of all contiguous trimer subsequences of the peptide; the surface residues predicted to comprise the binding site were then mapped over to the target in the bound complex for purposes of evaluating the accuracy of the prediction.  The ROC plots for these predictions are shown in Figure 33 and Figure 34 (red lines) and the TPR, FPR, PPV, and specificity statistics are shown in Table 9. At the same operating point as for the bound forms (likelihood ratio cutoff of 3.0, red arrows), three of the four binding sites identified using the bound form of the target protein were identified at a TPR of 50 percent or above (50%, 62%, and 82%, respectively for PDB:1MIX/PDB:1TWB, PDB:5CHY/PDB:1U8T, and PDB:1STP/PDB:1VWG), and two others were predicted with TPR of 43% and 49%, respectively (PDB:1RY4/PDB:1X8S and PDB:2AAV/PDB:2BP3), albeit at the cost of a somewhat higher FPR (13%, 5%, 8%, 20%, and 9%). For one structure (PDB:5CHY/PDB:1U8T), surprisingly, using the unbound structure resulted in a substantially improved prediction (PPV of 66% vs. 35%, due to a reduction in false positives). Overall for the eight targets, for predictions based on the unbound structures as compared to those from the bound forms, the mean TPR declined slightly from 45% to 43%, the mean FPR rose slightly from 9% to 11%, and the mean PPV declined from 44% to 33%.

**The VSPM approach can be extended to take into account geometric constraints based on the known sequence relationship among peptide residues.**

For five of the eight test structures, the search for plausible combinations of adjacent individual trimer poses based on geometric constraints produced reasonably good predictions of the binding loci, with a mean PPV of 68 percent, TPR of 66 percent, and FPR of 6 percent for the average of the ten highest scoring solution sets for each complex (see Table 11; “selected solution” refers to the optimal solutions as selected by selection criteria in which overall energy and positioning in high likelihood ratio surface regions are weighted equally (scaled by length); “best solution in set” refers to ranking of all solutions satisfying the geometric constraints by PPV; RMSD’s are alpha carbon average RMS distances as compared to the bound peptide for the residues present in the solution; “Len” is the chain length (number of residues) of the solution set).

For three of the complexes (PDB:1X8S, PDB:1VWG, and PDB:1U8T), the solution selected by the algorithm was in reasonable register with the bound peptide, with RMS deviations of 2.80Ǻ, 7.76Ǻ, and 5.45Ǻ, respectively, between the Cα positions of the predicted solution and those of the bound peptide (Figure 41). For two of the structures, however (PDB:2FF6 and PDB:2IVZ) the success rate in predicting the surface residues belonging to the binding site was zero percent. See Figure 33 and Figure 34, for TPR / FPR values for each complex for solutions based on the bound structure and unbound structure (black dot and red dot, respectively).

The success rate appears to be limited by two factors. First, the simple selection criterion by which the solution sets were ranked sometimes performs poorly in selecting an optimally predictive solution from the set of pose combinations that satisfy the geometric constraints, as can be seen from the statistics for the highest PPV solutions in each set (see Table 11, right hand columns). For example, for one of the complexes (PDB:2FF6) for which the selected solution had a true positive rate of zero in predicting the surface residues comprising the binding site, a solution existed in the set of geometrically plausible solutions that would have had a TPR of 68 percent and a FPR of only 1 percent. Given a sufficiently large training set, it would likely be possible to train or evolve a recognizer that would perform much better at selecting the optimal solution from the output of the tree search. A second factor is that the set of solutions emerging from the tree search may not necessarily contain any very good solutions if the population of sampled poses does not contain poses approximately representative of the bound positions of the corresponding residues in the peptide. Thus, for one of the structures (PDB:2IVZ), even the best solution in the set of solutions conforming to the geometric constraints has a PPV of only 28 percent.  This problem is likely in part due to insufficient sampling; presumably, with a much larger sample population, or, better yet, with enumerative evaluation of the entire surface, the required poses would be present.

|  |  | Selected Solution |  |  |  |  | Best Solution In Set |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | **PPV** | **TPR** | **FPR** | **Len** | **RMS** | **PPV** | **TPR** | **FPR** | **Len** | **RMS** |
| **1MK9** | 0.18 | 0.29 | 0.15 | 10.2 | 19.32 | 0.73 | 0.64 | 0.03 | 5.2 | 8.98 |
|  | 0.17 | 0.28 | 0.16 | 10.0 | 19.00 | 0.8 | 0.67 | 0.02 | 5.0 | 6.96 |
| (1MIX) | 0.11 | 0.05 | 0.02 | 5.0 |  | 0.4 | 0.11 | 0.00 | 5.0 |  |
| **1TWB** | 0.68 | 0.63 | 0.06 | 5.0 | 10.58 | 0.78 | 0.78 | 0.05 | 5.4 | 7.98 |
|  | 0.71 | 0.63 | 0.05 | 5.0 | 10.49 | 0.88 | 0.88 | 0.03 | 5.0 | 5.86 |
| (1ZSZ) | 0.64 | 0.56 | 0.07 | 6.0 |  | 0.93 | 0.88 | 0.01 | 6.0 |  |
| **1U8T** | 0.71 | 0.66 | 0.04 | 5.0 | 5.20 | 0.77 | 0.67 | 0.03 | 5.0 | 5.72 |
|  | 0.71 | 0.71 | 0.04 | 5.0 | 5.45 | 0.82 | 0.64 | 0.02 | 5.0 | 6.87 |
| (5CHY) | 0.34 | 0.71 | 0.23 | 13.0 |  | 0.57 | 0.29 | 0.04 | 5.0 |  |
| **1VWG** | 0.73 | 0.95 | 0.06 | 5.4 | 6.66 | 0.76 | 0.97 | 0.05 | 5.0 | 6.39 |
|  | 0.82 | 0.93 | 0.03 | 5.0 | 7.76 | 0.82 | 0.93 | 0.03 | 5.0 | 7.76 |
| (1STP) | 0.67 | 0.67 | 0.05 | 5.0 |  | 0.67 | 0.67 | 0.05 | 5.0 |  |
| **1X8S** | 0.68 | 0.57 | 0.06 | 5.8 | 3.79 | 0.84 | 0.68 | 0.03 | 5.8 | 3.79 |
|  | 0.60 | 0.56 | 0.08 | 6.0 | 2.80 | 0.86 | 0.75 | 0.03 | 6.0 | 2.80 |
| (1RY4) | 0.21 | 0.19 | 0.17 | 7.0 |  | 0.64 | 0.56 | 0.08 | 6.0 |  |
| **2BP3** | 0.57 | 0.49 | 0.11 | 5.4 | 11.16 | 0.74 | 0.72 | 0.07 | 6.0 | 12.28 |
|  | 0.13 | 0.11 | 0.20 | 5.0 | 13.67 | 0.83 | 0.56 | 0.03 | 5.0 | 8.76 |
| (2AAV) | 0.75 | 0.50 | 0.05 | 5.0 |  | 0.86 | 0.67 | 0.03 | 5.0 |  |
| **2FF6** | 0.00 | 0.00 | 0.05 | 11.0 | 14.84 | 0.86 | 0.67 | 0.01 | 6.0 | 14.04 |
|  | 0.00 | 0.00 | 0.04 | 11.0 | 15.07 | 0.87 | 0.68 | 0.01 | 6.0 | 14.18 |
| (1IJJ) | 0.00 | 0.00 | 0.07 | 5.0 |  | 0.68 | 0.68 | 0.02 | 5.0 |  |
| **2IVZ** | 0.00 | 0.00 | 0.09 | 5.0 | 23.06 | 0.28 | 0.48 | 0.09 | 8.0 | 15.73 |
|  | 0.00 | 0.00 | 0.09 | 5.0 | 22.94 | 0.28 | 0.48 | 0.09 | 8.0 | 15.61 |
| (1C5K) | 0.19 | 0.20 | 0.07 | 5.0 |  | 0.31 | 0.48 | 0.09 | 9.0 |  |
| **Averages for all** |  |  |  |  |  |  |  |  |  |  |
| **Best** | **0.45** | **0.45** | **0.08** |  | **11.83** | **0.72** | **0.70** | **0.04** |  | **9.36** |
| **Best** | **0.39** | **0.40** | **0.09** |  | **12.15** | **0.77** | **0.70** | **0.03** |  | **8.60** |
| **Unbnd** | **0.37** | **0.36** | **0.09** |  |  | **0.63** | **0.54** | **0.04** |  |  |
| **Averages for 1TWB, 1U8T, 1VWG, 1X8S, and 2BP3** |  |  |  |  |  |  |  |  |  |  |
| **Best** | **0.68** | **0.66** | **0.06** |  | **7.48** | **0.78** | **0.76** | **0.04** |  | **7.23** |
| **Best** | **0.60** | **0.59** | **0.08** |  | **8.03** | **0.84** | **0.75** | **0.03** |  | **6.41** |
| **Unbnd** | **0.52** | **0.53** | **0.11** |  |  | **0.73** | **0.61** | **0.04** |  |  |

Table 11. Statistics for VSPM predictions incorporating geometric constraints. Results from tree search for sets of individual poses conforming to geometric constraints. For each complex, the upper line reflects the average of the ten highest ranked solutions, the second line represents the highest ranked solution, and the third line represents the highest ranked solution based on the unbound structure.

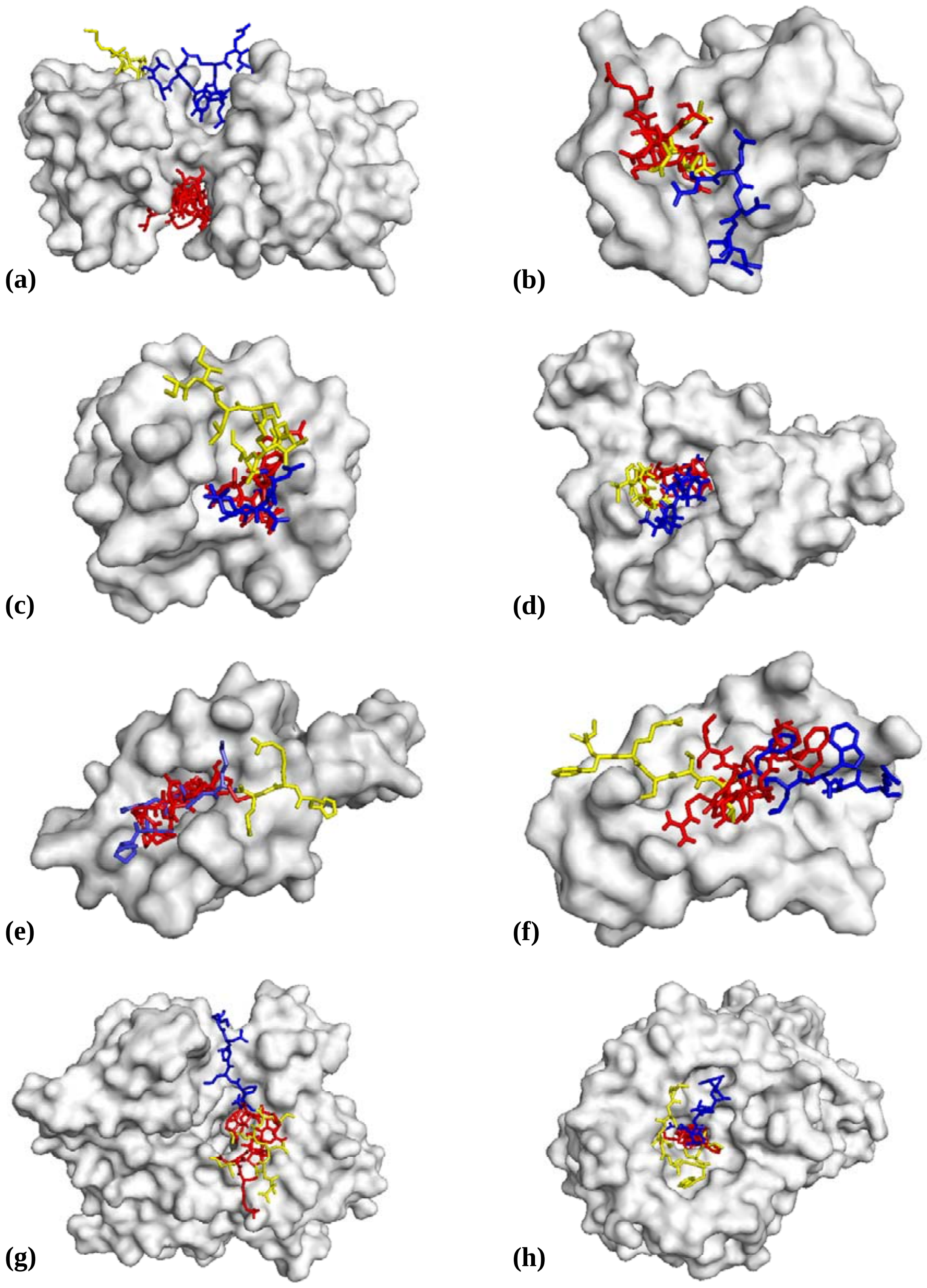
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

Figure 41 (Preceding page). VSPM solution sets with geometric constraints. Highest ranking solution sets of overlapping trimer poses satisfying geometric constraints for (a) PDB:1MK9, (b) PDB:1TWB, (c) PDB:1U8T, (d) PDB:1VWG, (e) PDB:1X8S, (f) PDB:2BP3, (g) PDB:2FF6, (h) PDB:2IVZ. Red: trimer poses belonging to solution set. Note that the solution set typically does not cover the entire peptide sequence. Blue: residues of bound peptide corresponding to residues present in solution set. Yellow: residues of bound peptide not corresponding to residues present in solution set.

### Algorithm performance generally, failure cases, and potential improvements

The essential concept underlying the VSPM strategy is the use of an arbitrary chemical moiety as a probe to map the properties of a surface by repeatedly positioning the probe at multiple points on the surface and computing the forces on the probe, the interaction energy, or some other quantity of interest. The results presented above show that this approach is capable of contributing useful information to the analysis of peptide-protein interactions; in particular, as applied to the eight test structures, it correctly predicts binding regions in three cases where the other prediction algorithms used for comparison failed. Although the computation time required is substantial – approximately 50 to 200 processor-hours per structure – the ability to analyze arbitrary or novel ligands without any need for bioinformatic data or data on evolutionary relationships may prove advantageous in appropriate cases, particularly where non-natural residues and/or non-biological sequences are involved.

The binding site prediction strategy performed poorly on two of the eight test complexes. In both cases, the reason appears to be that the probes, being small, are able to bind in topologically restricted locations – a deep crevice, in the case of PDB:1MK9, and a pore, in the case of PDB:2IVZ – that a full length peptide would be unable to penetrate. Both regions contain multiple loci for which individual trimer probes appeared to have quite high affinity, as often seems to be the case with pronounced surface concavities, and in the context of a Boltzmann weighting scheme, these loci swamp the signal from the remainder of the surface. If the full length 16-mer peptide of PDB:2IVZ somehow managed to work its way past the steep energy barrier and into the predicted binding region inside the pore, no doubt it would be very stably bound there, and likewise for the 14-mer peptide of PDB:1MK9 in the deep cleft. It should be possible to filter the mapping results to eliminate or downweight topologically restricted loci, perhaps via a local density constraint, but for purposes of this evaluation it was thought preferable to avoid such departures from generality.

In ideal terms, a virtual surface interrogation strategy might map interaction energy over a multidimensional space, as a function of both the spatial position on the target surface and additional dimensions to represent the orientation of the probe and the allowed bond rotations. For entities such as peptide trimers, which typically have approximately ten conformational degrees of freedom, such a space is too large to enumerate and evaluate exhaustively at any resolution likely to be useful, so it is necessary to reduce the set of points to be evaluated. This might perhaps be done most advantageously by specifying an arbitrary grid of target surface points and, at each point, using a suitable optimization procedure such as a genetic algorithm to determine a conformation and rotation of the probe that produces an energy value at or near the minimum achievable at that point. The local optimization would, in effect, simulate the effect of approaching a protein surface locus with a physical entity such as an AFM probe to which is conjugated (say) a peptide trimer: presumably the peptide trimer would arrange its bond rotations so as to achieve a local energy minimum. Future development plans include writing software to implement a strategy of this kind, but for purposes of prototyping and evaluating the potential utility of VSPM, it seemed preferable to leverage the already existing and well-tested functionality offered by AutoDock. The most laborious aspect of the computation required for energy mapping – the evaluation of the energy of a ligand at a particular position and in a particular conformation – is a task for which AutoDock is well suited. The sampling strategy worked adequately, but did require extensive reweighting to remove sources of bias that would not be present if the surface space were searched enumeratively rather than sampled.

It would also be possible, in an enumerative algorithm of the kind just described, to improve search efficiency and also generate data pre-filtered for a specific task, by incorporating conformational or other geometric constraints. For example, for a peptidic probe, one could sample and cluster conformations from a molecular dynamics trajectory of the probe, and use the resulting conformations as starting points for the local optimizations. Another constraint that would likely be useful in the context of peptide binding site prediction would be to prohibit poses in which either terminus of the probe is oriented toward the target surface in such a way as to prevent extension of the chain. Doing so would automatically eliminate a large number of solution poses that could not be part of a longer chain. Enumeration offers the further advantage of producing a single, deterministic solution, as compared to the sampling strategy employed here, where, owing to the stochastic nature of the genetic algorithm employed by AutoDock, results of successive runs may vary somewhat even for identical inputs, .

In the experiments presented here, no attempt was made to take into account unstructured regions of either the peptide or the protein target. In the case of unstructured regions of the protein, it is not feasible to do so since atomic coordinates are required for the energy evaluation. In the VSPM mapping, it would be possible to include probes representative of the regions of the peptide for which no coordinates are given in the PDB structure, but this was not done.

### Implications regarding mechanism of peptide-protein binding

It is surprising that a strategy that relies on sampling from a large number of (presumably) suboptimal binding poses involving small pieces of a bound peptide could provide any useful information at all, and even more surprising that it can do so on the basis of an unbound form of the target structure which differs significantly from the bound form. Of the hundreds of trimer docking poses comprising the sample set from which each map is constructed, typically very few, if any, match the position of the peptide in the bound complex closely enough so that the specific atom-level interactions between the probe and the target surface would be the same as those present in the bound complex depicted by the PDB structure, or even between the same pairs of atoms.  Those few poses that are in even approximate register with the corresponding part of the bound peptide are essentially never those with lowest estimated energy. Clearly, one thing that this prediction method is *not* doing is detecting interactions that are both essential for binding and highly sensitive to position.

This conclusion is underscored by the difficulty encountered in attempting to estimate individual residue positions by searching for combinations of individual poses from sample sets that satisfy geometric constraints such that they could plausibly represent the approximate structure of a bound peptide.  Even in the four out of eight cases (see Figure 33, black dots) where the search resulted in significantly improved identification of the surface residues involved in the binding site, the specific binding poses comprising the solution were typically in poor register with the position of the peptide in the PDB structure of the bound complex, and were sometimes oriented oppositely or at least differently, as the relatively high RMS distances testify (see Table 4). Further, the sample sets often did not contain poses from which a single set of overlapping poses could be constructed that would be in low-RMSD congruence with the bound position. If the position of the peptide as depicted in the PDB structure does represent the global energy minimum, then it may be that what the search algorithm is finding is various combinations of individual poses that correspond to local minima that may be nearby in configuration space and may represent fragments of possible transition states that lie between some initial encounter complex and the final stable state. It is also possible that the positions given in the PDB structures are artificially stabilized as a result of crystallization, and do not accurately reflect the dynamic nature of the complex as it exists in solution.

Several experimental observations arguably support the implication that a less fastidious, more probabilistic binding model is needed. In contrast to the near binary specificity observed in nucleic acid microarray experiments, a typical pure protein target will show significant affinity for a relatively large proportion of random 20-mer peptides. Experiments by others have shown that the affinity of a selected random sequence peptide for a specific protein target can be optimized by systematic substitution of a few individual residues, and that the binding energy improvements resulting from these substitutions can be approximately additive [33]. These and other observations, discussed in detail in Chapter 2, strongly suggest that these intermediate length peptides typically exist in solution (and on a peptide microarray surface) in a distribution of conformations, rather than in any stable, folded structure.

The hypothesis that seems most consistent with the computational results just presented, and with the experimental observations described in Chapters 2 and 3, is that the conformational and positional space surrounding a typical peptide binding site must contain many energy minima, and there must be many admissible transitions among those minima. In effect, the energy minima corresponding to peptide binding sites must usually lie in gentle valleys, not vertical bore-holes, in the positional and conformational energy landscape. As discussed in the preceding chapters, it is hypothesized that the interactions of interest are unlikely to be characterized by the usual “lock and key” paradigm of molecular docking, and that interactions between proteins and random peptides in the 8- to 20-mer size range are best understood via a model in which a given peptide is viewed as a relatively flexible entity made up of a series of moieties capable of energetically modest and non-fastidious interaction with suitable protein surface features, perhaps interspersed by relatively inert regions that contribute relatively little to $\Delta G$. Similarly, the protein may be thought of as exposing a mostly inert surface except for a scattering of small regions capable of interacting more strongly with the interactive moieties exposed by the peptide.

Such a hypothesis would explain the otherwise surprising success of the VSPM binding site predictions, and in particular those based on the unbound protein structures. A well-known limitation of computational docking is that it depends on estimating interatomic forces that are very sensitively dependent on distance, so that very small errors in the positions of the atoms in the target protein can result in large errors in the energies of the docking poses being searched. Given that structural models derived from X-ray or NMR experiments can, at best, provide only an approximate representation of one or a few conformational samples, “rigid” docking strategies fail to provide meaningful predictions in many cases. This is especially so for interactions wherein the protein surface shape changes appreciably upon binding of the ligand. If, however, the hypothesis is correct that the binding of intermediate-length peptides is less a matter of exact atom-level positioning and more a matter of finding bound positions capable of accommodating a few relatively non-fastidious interactions, then the problem of binding site prediction becomes much less dependent on perfectly accurate positioning of the protein surface atoms. Therefore, it should be possible to obtain reasonable estimates of binding loci using unbound target protein structures whose exact surface geometries may differ significantly from the final bound structure.

This binding model, if correct, also has intriguing implications for the design of peptide ligands of high affinity and specificity: it means that the goal should not be to find a single bound position and conformation of very high energy; instead, it should be to find a cluster of positions and conformations, which can be of lower energy, but which occupy points that are near each other in positional and conformational space and are not separated by high energy barriers. It may be noted that the latter goal is one that is potentially quite amenable to computational search and optimization, while the former is not.

## Conclusions

A method, virtual scanning probe mapping, has been introduced for assessing and spatially mapping the interactive properties of a surface with respect to an arbitrary chemical probe entity, by computational means. It has been shown that the method is capable of extracting information useful in predicting peptide binding loci on proteins. Unlike bioinformatic approaches, the method relies entirely on physics-based analysis, and so is not restricted to naturally occurring entities or to sequences that are well-represented in databases. As applied to a test set of eight peptide-protein complexes selected randomly from the PDB, the method performs well in predicting the protein surface residues belonging to the peptide binding site.

# CHAPTER 5: PREDICTION OF BINDING LOCI ON AKT-1 PROTEIN

The VSPM method described in Chapter 4 was used to analyze the binding properties of AKT-1 protein, a target for which a bivalent synbody was discovered experimentally by microarray screening (by others) [32]. Here a comparison is presented between the binding loci predicted by the VSPM algorithm and the AKT-1 surface moieties identified in cross-linking experiments (also by others).

## AKT structure and function

AKT, also called protein kinase B, is a serine/threonine protein kinase that is activated following ligand binding to G protein coupled receptors, receptor tyrosine kinases, or other cell surface receptors, whereupon AKT phosphorylates a large number of other signalling molecules within the cell [247-249]. Because of its involvement in signalling pathways that affect cell proliferation and suppression of apoptosis, AKT has received much attention as a possible cancer drug target [247, 250-252].

AKT has three distinct domains. The 118 residues (in AKT-1) beginning at the N-terminus correspond to the pleckstrin homology (PH) domain that binds phosphatidylinositol (3,4,5)-trisphosphate (PI(3,4,5)$P_3$). This is linked by a 39 amino acid hinge region of unknown structure [253] to the central catalytic domain, followed by an unstructured C-terminal hydrophobic domain [254]. In humans, AKT exists in three similar isoforms (AKT-1, AKT-2, and AKT-3); these have about 60% sequence identity in the PH domains, 25% in the hinge region, and 85% homology in the kinase domains. Discovery of a highly AKT-specific inhibitor has proved challenging, in part because the important residues in the two most easily targeted sites, the ATP binding pocket and the PH domains that bind the PI(3,4,5)$P_3$ head groups, are highly conserved, making it difficult to find inhibitors that can differentiate among related kinases [253].

Two x-ray structures (PDB:3CQU and PDB:3CQW) of the catalytic domain (residues 144 through 478 as numbered in PDB:3CQU) were available at the time the analysis reported here was performed [255]. These structures are co-crystals of AKT-1 with a pyrrolopyrimidine inhibitor and a synthetic peptide, “Crosstide”, which binds in the catalytic cleft of AKT. (Crosstide is a commercially available [256] peptide whose sequence (GRPRTSSFAEG) is derived from glycogen synthase kinase, is readily phosphorylated by AKT and other serine/threonine kinases, and is therefore used in kinase activity assays [257].)  Comparison of these PDB structures by backbone structural alignment shows them to be approximately identical except in the position of the loop around residue 160, which appears to control access to the active site, and is in a more “open” position in 3CQW by approximately 4Ǻ. Even leaving aside the (presumably) unstructured hinge region, AKT has been described as a “very flexible enzyme” [253]. Two additional similar x-ray structures, PDB:3MV5 and PDB:3MVH, again of the AKT-1 kinase domain co-crystallized with pyrimidine-based inhibitors, were recently published in May, 2010 [258]. An x-ray structure has also been published of the PH domain (PDB:1H10) [259], in complex with

$PI(3,4,5)P_3$. Another study describes a structure localizing a peptide inhibitor on the PH domain using NMR techniques [254].

### Cross-linking of synbody peptides to AKT

Cross-linking experiments were performed by Dr. Matthew P. Greving in which each of the two peptide binding elements of a synbody having high affinity and specificity for AKT-1 was bound and cross-linked to AKT-1, a trypsin digest was performed, and the resulting fragments were identified by mass spectrographic analysis. The peptide sequences were:

TRF23: FRGWAHIFFGPHVIYRGGSC

TRF26: AHKVVPQRQIRHAYNRYGSC

Table 12 shows the tryptic fragments identified on cross-linking, with the plausible cross-linked residues shown in bold type.  C-terminal lysines of fragments are assumed not to represent possible cross-link sites since trypsin would be unlikely to cleave at a site where a cross-link is present. Lysines whose amine groups are not surface-exposed in the x-ray structure are likewise assumed not to represent valid cross-link sites.  The cross-link site for fragment 5 cross-linked to TRF23 could not be determined because there is a 17-residue gap in the x-ray structure from residue M446, and the fragment cannot be matched to the structure. All fragments involving cross-links to peptide 23 contained only the N-terminal FR residues from the peptide. All fragments involving cross-links to peptide TRF26 contained the N-terminal fragment AHKVVPQR of the peptide.

| **Peptide** | **AKT Fragment** | **AKT Fragment Sequence** | **Possible cross-linked residues** |
|---|---|---|---|
| TRF23 | 1 | (R)DL**K**LENLMLDK(D)-Ox | K276 |
| TRF23 | 2 | (R)YYAMKIL**K**KEVIVAK(D) | K179, K182, K183 |
| TRF23 | 3 | (R)YYAMKIL**K**KEVIVAK(D)-Ox | K179, K182, K183 |
| TRF23 | 4 | (R)DL**K**LENLMLD**K**DGHIK(I) | K276, K284 |
| TRF23 | 5 | (R)RPHFPQFSYSASGTAKGDP(-) | None |
| TRF26 | 1 | (K)LLG**K**GTFG**K**VILVK(E) | K158, K163 |
| TRF26 | 2 | (R)FFAGIVWQHVYE**K**K(L) | K419 |
| TRF26 | 3 | (K)NVVYRDL**K**LENLMLDK(D) | K276 |

Table 12. Tryptic fragments from cross-linking peptides to AKT.

Based upon the data in Table 12, the following conclusions may be drawn:

1. There is an unambiguous cross-link of peptide TRF23 to AKT residue K276.
2. There is a cross-link of peptide TRF23 to at least one of AKT residues K179, K182, or K183; however, of these, only K182 has a reasonably surface-accessible amine.
3. A cross-link of peptide TRF23 to AKT residue K284 is possible, but not necessary to explain the data since the same fragment contains K276.
4. There are unambiguous cross-links of peptide TRF26 to AKT residues K276 and K419.
5. Peptide TRF26 cross-linked to AKT residue K158, K163, or both.

### Comparison of cross-link sites to binding sites predicted by VSPM

Figure 42(a) shows the two amines corresponding to AKT residues K182 (green) and K276 (blue). The region within a 1 nm radius of each is shaded in grey, indicating the approximate length of the cross-linker. Figure 42(b) shows the reverse face of the AKT molecule, with K284 shown in cyan and the surface locus corresponding to K183 (almost completely buried) in green. Figure 42(c) and (d) show a spatial mapping of the binding likelihood ratio of peptide 23 as computed by the VSPM method described in Chapter 4 (darker red indicates higher likelihood). Figure 42(e) and (f) show a spatial mapping of the binding likelihood ratio of the N-terminal trimer (FRG) of peptide 23, that being the point at which the peptide would be cross-linked. As is apparent from Figure 42(c) and (e), the VSPM prediction is in complete agreement with the cross-linking data regarding the region surrounding residue K276 (blue in Figure 42(a), (c), and (e)). The predicted binding propensity map is less satisfying with respect to residue K182 (green in Figure 42(a), (c), and (e)), but given the size of the cross-linker it is plausible that a peptide bound where indicated by the higher probability region could cross-link to that residue. The VSPM map also indicates a high-probability region on the opposite face (Figure 42 (d)), but it appears that the binding propensity at that site is not due to the N-terminal portion of the peptide (Figure 42(f)).

These data are consistent with the known peptide binding properties of AKT-1. Residue K276 (blue in Figure 42 (a), (c), and (e)) lies inside a furrow that can be seen running vertically in the figure; the Crosstide peptide, which mimics a natural substrate of AKT, binds in the same furrow, as seen in  Figure 43, which shows the x-ray structure of PDB:3CQU with the Crosstide peptide bound.

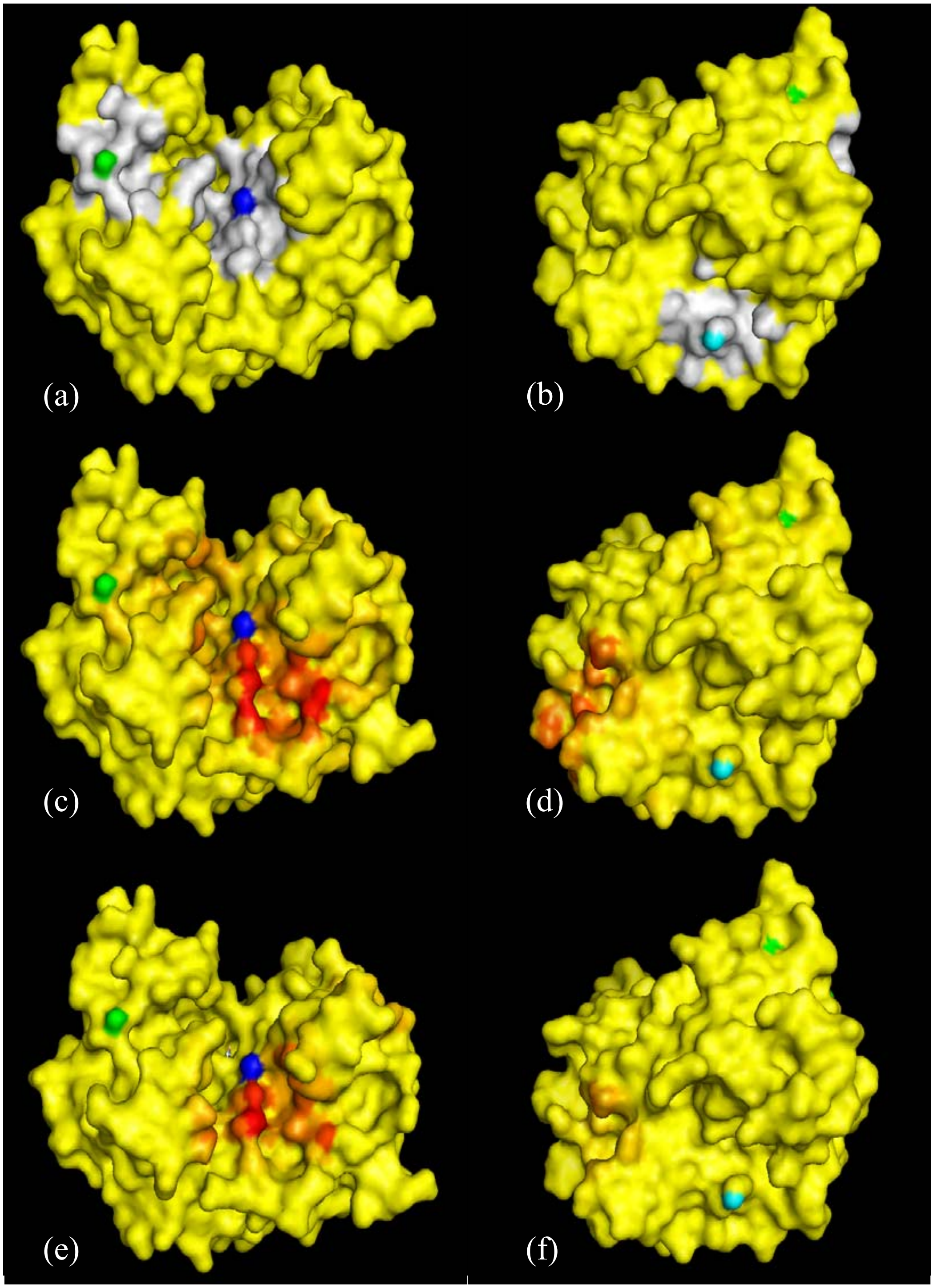

Figure 42. Cross-link sites and predicted binding sites, peptide TRF23

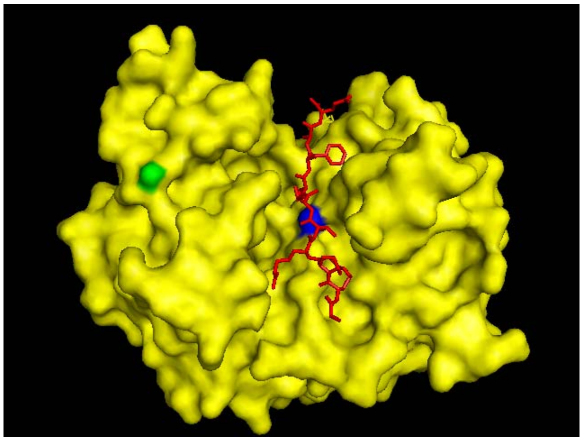

Figure 43. AKT-1 showing position of Crosstide peptide.

The cross-linking data indicates that peptide TRF26 is cross-linking to at least three distinct AKT residues. .
Figure 44 shows the locations of the amines on the AKT surface: K276 (blue, Figure 44(a), (c), and (e)), K158 and K163 (green, Figure 44(a), (c), and (e)), and K419 (cyan,

Figure 44(b), (d), and (f)). Again, the VSPM algorithm finds the highest binding likelihood in the substrate binding cleft, in the region surrounding residue K276, and to a lesser extent on the opposite face within cross-linking distance from residue K419.

### Implications regarding binding specificity

These data provide further support for the hypothesis developed in the preceding chapters that intermediate length random peptides are unlikely to bind in a single fixed configuration. The N-terminal amine of peptide TRF23 cross-links to at least two distinct loci that are ~22Ǻ apart, and the N-terminus of peptide TRF26 cross-links to at least three distinct loci, none of which is closer than ~17Ǻ from any of the others, and one of which is on the opposite face of the

AKT molecule from the others. Clearly, the cross-linking data cannot be explained by a fixed position of either peptide. The VSPM predictions also indicate two distinct binding regions for each peptide, and appear to be generally consistent with the cross-linking data.

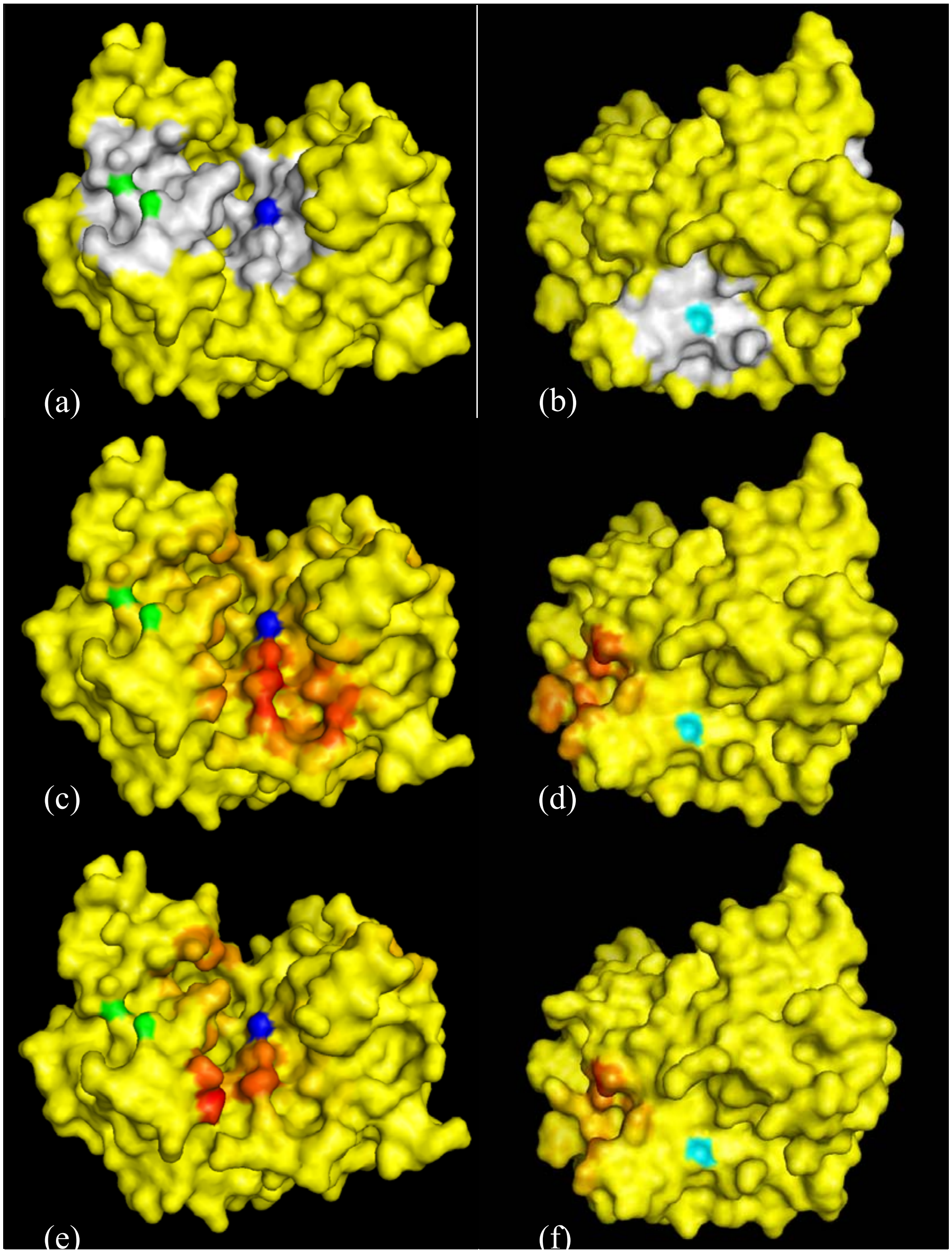

Figure 44. Cross-link sites and predicted binding sites, peptide TRF26.

# CHAPTER 6: CONCLUDING REMARKS

The central argument of the preceding chapters is that peptide-protein interfaces are dynamic entities, and that the binding behavior of random, intermediate length peptides must be evaluated in probabilistic terms that take into account the multiplicity of positions and conformations that may exist in an ensemble of bound complexes, rather than in terms of a single bound position in a single binding site.

The peptides of interest are flexible and adopt a broad distribution of conformations in their free solvated state. Evidence for this conclusion, discussed in Chapters 2 and 3, includes modeling studies of the range of conformational variability resulting from bond rotations; statistical data from a large number of microarray experiments; conformational sampling and clustering from molecular dynamics modeling, and theoretical analysis described in the pertinent scientific literature.

A simple "lock and key" model fails to explain the observed kinetics of peptide-protein binding. The promiscuity of peptides in microarray and SPR experiments (Chapter 2), the long incubation times required for equilibrium to be established (Chapter 2), the positional variability observed in PDB structures where multiple models of the same interface are available (Chapter 3), the surprising success of a probabilistically driven binding site prediction model (Chapter 4), and experimental evidence indicating multiple cross-linking loci (Chapter 5) are difficult to reconcile with an assumption that interfaces are static.

Conversely, these observations, as well as the observed binding behavior in SPR experiments (Chapter 2) are readily explained if it is assumed that multiple bound configurations are present.

In the course of developing this argument, several inquiries were pursued that yielded results of possible future utility for the purpose of discovering and optimizing peptide ligands for use in synbodies. As described in Chapter 3, a comprehensive dataset was constructed, containing all peptide-protein interfaces present in the Protein Data Bank at the time the dataset was compiled that could be reasonably regarded as comprising a protein and a surface-bound intermediate-length peptide. The interfaces were extracted, with anomalous or non-standard representations corrected, and were energy minimized and analyzed to extract a wealth of statistical data about interface geometry and other measurable characteristics. Non-bonded interactions (hydrogen bonds, salt bridges, cation-pi interactions, and hydrophobic interactions) were identified and described in detail in a relational database, and energies were evaluated according to a detailed model, with weights trained on the basis of experimentally determined affinities. These data provide a basis for inferring heuristics, discussed in detail in Chapter 3, to inform the design and selection of peptide ligands. The “sanitized” and minimized interface dataset and the relational database containing the results of analysis have been made publicly available via a web repository, together with the source code for the PopTop and VSPM software (see Appendix 1 for locations).

Also developed in the course of the inquiry was a novel general technique for spatial mapping of the chemical or other characteristics of a target macromolecule by interrogating its surface with small probe entities whose interactions provide a local measure of the interactive characteristics desired to be mapped. In Chapter 4, this "Virtual Scanning Probe Mapping" (VSPM) technique was applied to produce a spatial map of the likelihood of binding by a specified peptide, and was shown to predict peptide binding sites in a test set of eight PDB interfaces more accurately than other published methods. The method was also shown, in Chapter 5, to predict the likely binding region of two specific synbody peptides on AKT-1 protein in a manner reasonably consistent with evidence from cross-linking experiments and with the known peptide-binding behavior of AKT-1.

By way of indicating future directions, more extensive testing against specific experimental data would be useful. For example, the VSPM strategy can be readily extended to rank peptides in terms of likely binding affinity for a specified protein target, and the resulting predictions can be evaluated in comparison with microarray or SPR binding patterns, given the availability of suitable experimental data, extraction of which is currently underway. Efforts are also underway, in a collaboration between the Center for Innovations in Medicine and another laboratory, to obtain an x-ray structure of a synbody bound to a protein target. If obtainable, data of that kind would be of great interest for validating VSPM binding site predictions (again with some reservations about bias due to crystallization). For another example, the hypotheses of Chapter 2 regarding binding kinetics and equilibration times should be readily amenable to falsification given appropriate experimental data.

There are also obvious improvements to be made in the implementation of the VSPM algorithm. As noted in Chapter 3, an enumerative search strategy is likely to yield better results than the sampling approach used in the existing prototype. And pre-determining the ensemble of probe conformations on the basis of molecular dynamics sampling and clustering of the peptide's conformations in solution would likely improve accuracy and computational efficiency. An improved VSPM algorithm, combined with a classifier based on the heuristics derived in Chapter 3, might prove capable of substantially improving peptide microarray screening efficiency by filtering out peptides that are predictably unlikely to bind.

From an engineering design standpoint, peptide ligands offer both challenge and opportunity. The challenge arises from the complex and dynamic binding behavior, arguably requiring a model that takes into account many possible intermediate states and reaction paths. The opportunity lies in the possibility of devising ligands that can recognize complex macromolecular targets "holistically", in a manner that is not necessarily dependent on the interaction of a specific moiety at a specific site. For that potential to be realized, considerable work remains to be done, both from a theoretical standpoint and in development of practicable models and algorithms.

## REFERENCES

1. Gronwall C, Stahl S: **Engineered affinity proteins-Generation and applications**. *Journal of Biotechnology* 2009, **140**(3-4):254-269.

2. Hey T, Fiedler E, Rudolph R, Fiedler M: **Artificial, non-antibody binding proteins for pharmaceutical and industrial applications**. *Trends in Biotechnology* 2005, **23**(10):514-522.

3. Skerra A: **Alternative non-antibody scaffolds for molecular recognition**. *Current Opinion in Biotechnology* 2007, **18**(4):295-304.

4. Combes RD: **New measures on animal experimentation in the UK will improve animal welfare and scientific research**. *Atla-Alternatives to Laboratory Animals* 1999, **27**(3):309-316.

5. Qiu XQ, Wang H, Cai B, Wang LL, Yue ST: **Small antibody mimetics comprising two complementarity-determining regions and a framework region for tumor targeting**. *Nature Biotechnology* 2007, **25**(8):921-929.

6. Klingbeil C, Hsu DH: **Pharmacology and safety assessment of humanized monoclonal antibodies for therapeutic use**. *Toxicologic Pathology* 1999, **27**(1):1-3.

7. Pelech S: **Tracking cell signaling protein expression and phosphorylation by innovative proteomic solutions**. *Current Pharmaceutical Biotechnology* 2004, **5**(1):69-77.

8. Nilsson FY, Tolmachev V: **Affibody (R) molecules: New protein domains for molecular imaging and targeted tumor therapy**. *Current Opinion in Drug Discovery & Development* 2007, **10**(2):167-175.

9. Huang J, Koide A, Nettle KW, Greene GL, Koide S: **Conformation-specific affinity purification of proteins using engineered binding proteins: Application to the estrogen receptor**. *Protein Expression and Purification* 2006, **47**(2):348-354.

10. Gunneriusson E, Nord K, Uhlen M, Nygren PA: **Affinity maturation of a Taq DNA polymerase specific affibody by helix shuffling**. *Protein Engineering* 1999, **12**(10):873-878.

11. Nord K, Gunneriusson E, Uhlen M, Nygren PA: **Ligands selected from combinatorial libraries of protein A for use in affinity capture of**

**apolipoprotein A-1(M) and Taq DNA polymerase**. *Journal of Biotechnology* 2000, **80**(1):45-54.

12. Andersson M, Ronnmark J, Arestrom I, Nygren PA, Ahlborg N: **Inclusion of a non-immunoglobulin binding protein in two-site ELISA for quantification of human serum proteins without interference by heterophilic serum antibodies**. *Journal of Immunological Methods* 2003, **283**(1-2):225-234.

13. Kulkarni AA, Weiss AA, Iyer SS: **Glycan-Based High-Affinity Ligands for Toxins and Pathogen Receptors**. *Medicinal Research Reviews* 2010, **30**(2):327-393.

14. Friedman M, Stahl S: **Engineered affinity proteins for tumour-targeting applications**. *Biotechnology and Applied Biochemistry* 2009, **53**:1-29.

15. Audie J, Boyd C: **The Synergistic Use of Computation, Chemistry and Biology to Discover Novel Peptide-Based Drugs: The Time is Right**. *Current Pharmaceutical Design*, **16**(5):567-582.

16. **Peptide Drug Discovery Research Reenergized**. In: *Genetic Engineering News.* vol. 26; 2006.

17. Unal EB, Gursoy A, Erman B: **VitAL: Viterbi Algorithm for de novo Peptide Design**. *PLoS ONE*, **5**(6).

18. Wang ST, Feng JN, Guo JW, Li Y, Sun YX, Qin WS, Hu MR, Shen BF: **Structural-based rational design of an antagonist peptide that inhibits the ribosome-inactivating activity of ricin A chain**. *International Journal of Peptide Research and Therapeutics* 2005, **11**(3):211-218.

19. Fassina G, Cassani G, Corti A: **Binding of Human TUmor Necrosis Factor alpha to Multimeric Complementary Peptides**. *Arch Biochem Biophys* 1992, **296**(1):137-143.

20. Selz KA, Mandell AJ, Shlesinger MF, Arcuragi V, Owens MJ: **Designing human m(1) muscarinic receptor-targeted hydrophobic eigenmode matched peptides as functional modulators**. *Biophysical Journal* 2004, **86**(3):1308-1331.

21. Selz KA, Samoylova TI, Samoylov AM, Vodyanoy VJ, Mandell AJ: **Designing allosteric peptide ligands targeting a globular protein**. *Biopolymers* 2007, **85**(1):38-59.

22. Jackrel ME, Valverde R, Regan L: **Redesign of a protein–peptide interaction: Characterization and applications**. *Protein Science* 2009, **18**(4):762-774.

23. Su ZD, Vinogradova A, Koutychenko A, Tolkatchev D, Ni F: **Rational design and selection of bivalent peptide ligands of thrombin incorporating P-4-P-1 tetrapeptide sequences: from good substrates to potent inhibitors**. *Protein Engineering Design & Selection* 2004, **17**(8):647-657.

24. Shrivastava A, Nunn AD, Tweedle MF: **Designer Peptides: Learning from Nature**. *Current Pharmaceutical Design* 2009, **15**(6):675-681.

25. Liu FF, Wang T, Dong XY, Sun Y: **Rational design of affinity peptide ligand by flexible docking simulation**. *Journal of Chromatography A* 2007, **1146**(1):41-50.

26. Lataifeh A, Beheshti S, Kraatz HB: **Designer Peptides: Attempt to Control Peptide Structure by Exploiting Ferrocene as a Scaffold**. *European Journal of Inorganic Chemistry* 2009(22):3205-3218.

27. Audie J, Boyd C: **The Synergistic Use of Computation, Chemistry and Biology to Discover Novel Peptide-Based Drugs: The Time is Right**. *Current Pharmaceutical Design* 2009, **16**(5):567-582.

28. Johnston SA, Woodbury N, Diehnelt CW, Belcher P, Gupta N, Zhao Z-G, Greving M, Emery J: **Synthetic Antibodies**. In: *European Patent Office, Application No WO2009140039.* 2008.

29. Betanzos CM, Gonzalez-Moa M, Johnston SA, Svarovsky SA: **Facile labeling of lipoglycans with quantum dots**. *Biochemical and Biophysical Research Communications* 2009, **380**(1):1-4.

30. Boltz KW, Gonzalez-Moa MJ, Stafford P, Johnston SA, Svarovsky SA: **Peptide microarrays for carbohydrate recognition**. *Analyst* 2009, **134**(4):650-652.

31. Williams BAR, Diehnelt CW, Belcher P, Greving M, Woodbury NW, Johnston SA, Chaput JC: **Creating Protein Affinity Reagents by Combining Peptide Ligands on Synthetic DNA Scaffolds**. *Journal of the American Chemical Society* 2009, **131**(47):17233-17241.

32. Diehnelt C, et al.: **Discovery of High Affinity Protein Binding Ligands -- Backwards**. *PLOS One* 2010, **5**(5):e10728. doi:10710.11371/journal.pone.0010728.

33. Greving MP, Belcher P, Diehnelt C, Gonzalez-Moa, M., Emery J, Fu J, Johnston SA, Woodbury NW: **Thermodynamic additivity of sequence variations: An algorithm for creating High Affinity Peptides without large libraries or structural information**. *PLOS One* 2010, **5**(11):e15432.

34. Chun PW: **Molecular-level thermodynamic switch controls chemical equilibrium in sequence-specific hydrophobic interaction of 35 dipeptide pairs**. *Biophysical Journal* 2003, **84**(2):1352-1369.

35. Cui J, Han LY, Lin HH, Zhang HL, Tang ZQ, Zheng CJ, Cao ZW, Chen YZ: **Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties**. *Molecular Immunology* 2007, **44**(5):866-877.

36. Donnes P, Kohlbacher O: **SVMHC: a server for prediction of MHC-binding peptides**. *Nucleic Acids Research* 2006, **34**:W194-W197.

37. Doytchinova IA, Blythe MJ, Flower DR: **Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201**. *Journal of Proteome Research* 2002, **1**(3):263-272.

38. Fagerberg T, Cerottini JC, Michielin O: **Structural prediction of peptides bound to MHC class I**. *Journal of Molecular Biology* 2006, **356**(2):521-546.

39. Guan PP, Doytchinova IA, Walshe VA, Borrow P, Flower DR: **Analysis of peptide-protein binding using amino acid descriptors: Prediction and experimental verification for human histocompatibility complex HLA-A*0201**. *Journal of Medicinal Chemistry* 2005, **48**(23):7418-7425.

40. Hertz T, Yanover C: **PepDist: A new framework for protein-peptide binding prediction based on learning peptide distance functions**. *Bmc Bioinformatics* 2006, **7**.

41. Hoare SRJ: **Mechanisms of peptide and nonpeptide ligand binding to class B G-protein coupled receptors**. *Drug Discovery Today* 2005, **10**(6):417-427.

42. Ishida H, Vogel HJ: **Protein-peptide interaction studies demonstrate the versatility of calmodulin target protein binding**. *Protein and Peptide Letters* 2006, **13**(5):455-465.

43. Kasson PM, Rabinowitz JD, Schmitt L, Davis MM, McConnell HM: **Kinetics of peptide binding to the class II MHC protein I-E**. *Biochemistry* 2000, **39**(5):1048-1058.

44. Klepeis JL, Ierapetritou MG, Floudas CA: **Protein folding and peptide docking: A molecular modeling and global optimization approach**. *Computers & Chemical Engineering* 1998, **22**:S3-S10.

45. Weng GZ: **Exploring protein-protein interactions by peptide docking protocols**. In: *G Protein Pathways Part B: G Proteins and Their Regulators.* vol. 344; 2002: 577-586.

46. Yanover C, Hertz T: **Predicting protein-peptide binding affinity by learning peptide-peptide distance functions**. *Lecture notes in computer science* 2005, **3500**:456-471.

47. Zhang L, Shao C, Zheng DX, Gao YH: **An integrated machine learning system to computationally screen protein databases for protein binding peptide ligands**. *Molecular & Cellular Proteomics* 2006, **5**(7):1224-1232.

48. Zhao CY, Zhang HX, Luan F, Zhang RS, Liu MC, Hu ZD, Fan BT: **QSAR method for prediction of protein-peptide binding affinity: Application to MHC class I molecule HLA-A*0201**. *Journal of Molecular Graphics & Modelling* 2007, **26**(1):246-254.

49. Berezhkovskiy LM: **The analysis of peptide affinity and its binding kinetics to DR1DW1 major histocompatibility complex protein**. *Biophysical Chemistry* 1999, **77**(2-3):183-194.

50. Atkins PW: **Physical Chemistry, 3d Ed.**, Third edn. Oxford: Oxford University Press; 1986.

51. Dill KA, Bromberg S: **Molecular Driving Forces**. New York: Garland Science; 2003.

52. Luque I, Freire E: **Structure-based prediction of binding affinities and molecular design of peptide ligands**. In: *Energetics of Biological Macromolecules, Pt B.* vol. 295; 1998: 100-127.

53. Ajay, Murcko MA: **Computational methods to predict binding free energy in ligand-receptor complexes**. *Journal of Medicinal Chemistry* 1995, **38**(26):4953-4967.

54. Sotriffer CA, Sanschagrin P, Matter H, Klebe G: **SFCscore: Scoring functions for affinity prediction of protein-ligand complexes**. *Proteins-Structure Function and Bioinformatics* 2008, **73**(2):395-419.

55. Zhang C, Liu S, Zhu QQ, Zhou YQ: **A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes**. *Journal of Medicinal Chemistry* 2005, **48**(7):2325-2335.

56. Kleywegt GJ, Jones TA: **Phi/psi-chology: Ramachandran revisited**. *Structure* 1996, **4**(12):1395-1400.

57. Dunbrack RL, Karplus M: **Conformational Analysis Of The Backbone-Dependent Rotamer Preferences Of Protein Side-Chains**. *Nature Structural Biology* 1994, **1**(5):334-340.

58. Koch K, Zöllner F, Neumann S, Kummert F, Sagerer G: **Comparing bound and unbound protein structures using energy calculation and rotamer statistics**. *In Silico Biology* 2002, **2**:0032.

59. Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA: **The relationship of protein conservation and sequence length**. *BMC Evolutionary Biology* 2002, **2**(20 Cited December 10, 2002).

60. Liwo A, Czaplewski C, Oldziej S, Scheraga HA: **Computational techniques for efficient conformational sampling of proteins**. *Current Opinion in Structural Biology* 2008, **18**(2):134-139.

61. Flory PJ: **Statistical Mechanics of Chain Molecules**. New York: John Wiley & Sons; 1969.

62. Pappu RV, Wang X, Vitalis A, Crick SL: **A polymer physics perspective on driving forces and mechanisms for protein aggregation**. *Arch Biochem Biophys* 2008, **469**(1):132-141.

63. Dondos A: **Dependence of the theta temperature on the rigidity of the polymers and the quality of the solvents**. *Macromolecules* 1992, **25**(22):6069-6071.

64. Sun ST, Nishio I, Swislow G, Tanaka T: **The coil-globule transition -- radius of gyration of polystyrene in cyclohexane**. *Journal of Chemical Physics* 1980, **73**(12):5971-5975.

65. Blanco FJ, Rivas G, Serrano L: **A Short Linear Peptide That Folds Into A Native Stable Beta-Hairpin In Aqueous-Solution**. *Nature Structural Biology* 1994, **1**(9):584-590.

66. Searle MS, Williams DH, Packman LC: **A Short Linear Peptide Derived From The N-Terminal Sequence Of Ubiquitin Folds Into A Water-Stable Nonnative Beta-Hairpin**. *Nature Structural Biology* 1995, **2**(11):999-1006.

67. Marqusee S, Robbins VH, Baldwin RL: **Unusually Stable Helix Formation In Short Alanine-Based Peptides**. *Proceedings of the National Academy of Sciences of the United States of America* 1989, **86**(14):5286-5290.

68. Ho BK, Dill KA: **Folding very short peptides using molecular dynamics**. *PLOS Computational Biology* 2006, **2**(4):228-237.

69. Thomas A, Deshayes S, Decaffmeyer M, Van Eyck MH, Charloteaux B, Brasseur R: **Prediction of peptide structure: How far are we?** *Proteins-Structure Function and Bioinformatics* 2006, **65**(4):889-897.

70. Meszaros B, Simon I, Dosztanyi Z: **Prediction of Protein Binding Regions in Disordered Proteins**. *PLoS Computational Biology* 2009, **5**(5):e1000376.

71. Meszaros B, Tompa P, Simon I, Dosztanyi Z: **Molecular principles of the interactions of disordered proteins**. *Journal of Molecular Biology* 2007, **372**(2):549-561.

72. Tompa P: **Intrinsically unstructured proteins**. *Trends in Biochemical Sciences* 2002, **27**(10):527-533.

73. Dunker AK, Oldfield CJ, Meng JW, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, Uversky VN: **The unfoldomics decade: an update on intrinsically disordered proteins**. *Bmc Genomics* 2007, **9**.

74. Maupetit J, Derreumaux P, Tuffery P: **PEP-FOLD: an online resource for de novo peptide structure prediction**. *Nucleic Acids Research* 2009, **37**:W498-W503.

75. Zhang Y: **Protein structure prediction: when is it useful?** *Current Opinion in Structural Biology* 2009, **19**(2):145-155.

76. Cozzetto D, Tramontano A: **Advances and Pitfalls in Protein Structure Prediction**. *Current Protein & Peptide Science* 2008, **9**(6):567-577.

77. Nicosia G, Stracquadanio G: **Generalized Pattern Search Algorithm for Peptide Structure Prediction**. *Biophysical Journal* 2008, **95**(10):4988-4999.

78. Shi S, Pei J, Sadreyev R, Kinch L, Majumdar I, Tong J, Cheng H, Kim B, Grishin N: **Analysis of CASP8 targets, predictions and assessment methods**. In: *Database.* vol. 2009; 2009: doi:10.1093/database/bap1003.

79. Kaufmann KW, Lemmon GH, DeLuca SL, Sheehan JH, Meiler J: **Practically Useful: What the ROSETTA Protein Modeling Suite Can Do for You**. *Biochemistry* 2009, **49**(14):2987-2998.

80. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA: **The protein folding problem: when will it be solved?** *Current Opinion in Structural Biology* 2007, **17**(3):342-346.

81. Ozkan SB, Wu GA, Chodera JD, Dill KA: **Protein folding by zipping and assembly**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(29):11987-11992.

82. Snow CD, Nguyen N, Pande VS, Gruebele M: **Absolute comparison of simulated and experimental protein-folding dynamics**. *Nature* 2002, **420**(6911):102-106.

83. Subramanian R, Kazerounian K: **Improved molecular model of a peptide unit for proteins**. *Journal of Mechanical Design* 2007, **129**(11):1130-1136.

84. Bujnicki JM: **Protein-structure prediction by recombination of fragments**. *Chembiochem* 2006, **7**(1):19-27.

85. Kim DE, Chivian D, Baker D: **Protein structure prediction and analysis using the Robetta server**. *Nucleic Acids Research* 2004, **32**:W526-W531.

86. Ngan SC, Inouye MT, Samudrala R: **A knowledge-based scoring function based on residue triplets for protein structure prediction**. *Protein Engineering Design & Selection* 2006, **19**(5):187-193.

87. Heuser P, Wohlfahrt G, Schomburg D: **Efficient methods for filtering and ranking fragments for the prediction of structurally variable regions in proteins**. *Proteins-Structure Function and Genetics* 2004, **54**(3):583-595.

88. Simons KT, Kooperberg C, Huang E, Baker D: **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions**. *Journal of Molecular Biology* 1997, **268**(1):209-225.

89. Perczel A, Jakli I, Csizmadia IG: **Intrinsically stable secondary structure elements of proteins: A comprehensive study of folding units of proteins by computation and by analysis of data determined by x-ray crystallography**. *Chemistry-a European Journal* 2003, **9**(21):5332-5342.

90. Bystroff C, Simons KT, Han KF, Baker D: **Local sequence-structure correlations in proteins**. *Current Opinion in Biotechnology* 1996, **7**(4):417-421.

91. Lee MR, Tsai J, Baker D, Kollman PA: **Molecular dynamics in the endgame of protein structure prediction**. *Journal of Molecular Biology* 2001, **313**(2):417-430.

92. Fogolari F, Pieri L, Dovier A, Bortolussi L, Giugliarelli G, Corazza A, Esposito G, Viglino P: **Scoring predictive models using a reduced representation of proteins: model and energy definition**. *Bmc Structural Biology* 2007, **7**.

93. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices**. *Journal of Molecular Biology* 1999, **292**(2):195-202.

94. Pirovano W, Heringa J: **Protein secondary structure prediction**. *Methods in Mol BIo* 2010, **609**:327-348.

95. de Bakker PIW, Bateman A, Burke DF, Miguel RN, Mizuguchi K, Shi J, Shirai H, Blundell TL: **HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families**. *Bioinformatics* 2001, **17**(8):748-749.

96. Shi JY, Blundell TL, Mizuguchi K: **FUGUE: Sequence-structure homology recognition using environment-specific substitution tables**

**and structure-dependent gap penalties**. *Journal of Molecular Biology* 2001, **310**(1):243-257.

97. Etchebest C, Benros C, Hazout S, de Brevern AG: **A structural alphabet for local protein structures: Improved prediction methods**. *Proteins-Structure Function and Bioinformatics* 2005, **59**(4):810-827.

98. O'Donohue MF, Minasian E, Leach SJ, Burgess AW, Treutlein HR: **PEPCAT - A new tool for conformational analysis of peptides**. *Journal of Computational Chemistry* 2000, **21**(6):446-461.

99. Tanizaki S, Clifford J, Connelly BD, Feig M: **Conformational sampling of peptides in cellular environments**. *Biophysical Journal* 2008, **94**(3):747-759.

100. Knecht V, Mohwald H, Lipowsky R: **Conformational diversity of the fibrillogenic fusion peptide B18 in different environments from molecular dynamics simulations**. *Journal of Physical Chemistry B* 2007, **111**(16):4161-4170.

101. Voelz VA, Shell MS, Dill KA: **Predicting Peptide Structures in Native Proteins from Physical Simulations of Fragments**. *PLOS Computational Biology* 2009, **5**(2):e1000281.

102. Thomas A, Deshayes S, Decaffmeyer M, Van Eyck MH, Benoit CB, Brasseur R: **PepLook: An innovative in silico tool for determination of structure, polymorphism and stability of peptides**. *Peptides for Youth* 2009, **611**:459-460.

103. Abagyan R, Totrov M: **Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins**. *Journal of Molecular Biology* 1994, **235**(3):983-1002.

104. Morales LB, Garduno-Juarez R, Aguilar-Alvarado JM, Riveros-Castro FJ: **A parallel tabu search for conformational energy optimization of oligopeptides**. *Journal of Computational Chemistry* 2000, **21**(2):147-156.

105. Wang ZQ, Pachter R: **Prediction of peptide conformation: The adaptive simulated annealing approach**. *Journal of Computational Chemistry* 1997, **18**(3):323-329.

106. Daura X: **Molecular dynamics simulation of peptide folding**. *Theoretical Chemistry Accounts* 2006, **116**(1-3):297-306.

107. Derreumaux P: **Ab initio prediction of polypeptide structure from its sequence**. *Computer Physics Communications* 1999, **122**:139-140.

108. Derreumaux P: **Ab initio polypeptide structure prediction**. *Theoretical Chemistry Accounts* 2000, **104**(1):1-6.

109. Derreumaux P: **Insight into protein topology from Monte Carlo simulations**. *Journal of Chemical Physics* 2002, **117**(7):3499-3503.

110. Nakamura H: **Prediction of peptide conformation using a scale-transformed entropy sampling algorithm**. *Computational Biology and Chemistry* 2004, **28**(1):61-66.

111. **Flexweb: Analysis of Flexibility in Biomolecules and Networks** [http://flexweb.asu.edu]

112. Wells S, Menor S, Hespenheide B, Thorpe MF: **Constrained geometric simulation of diffusive motion in proteins**. *Physical Biology* 2005, **2**(4):S127-S136.

113. Maupetit J, Derreumaux P, Tuffery P: **A Fast Method for Large-Scale De Novo Peptide and Miniprotein Structure Prediction**. *Journal of Computational Chemistry* 2009, **31**(4):726-738.

114. **Robetta: Full-chain Protein Structure Prediction Server** [http://www.robetta.org/]

115. Fischer D: **Servers for protein structure prediction**. *Current Opinion in Structural Biology* 2006, **16**(2):178-182.

116. **PEPstr: PEPTIDE TERTIARY STRUCTURE PREDICTION SERVER** [http://www.imtech.res.in/raghava/pepstr/]

117. Kaur H, Garg A, Raghava GPS: **PEPstr: A de novo method for tertiary structure prediction of small bioactive peptides**. *Protein and Peptide Letters* 2007, **14**(7):626-631.

118. **Biosiris: Peplook Details** [http://www.biosiris.com/products-and-services/peplook.html]

119. Pandini A, Fornili A, Kleinjung J: **Structural alphabets derived from attractors in conformational space**. *Bmc Bioinformatics* 2010, **11**.

120. Camproux AC, Gautier R, Tuffery P: **A hidden Markov model derived structural alphabet for proteins**. *Journal of Molecular Biology* 2004, **339**(3):591-605.

121. Camproux AC, Tuffery P: **Hidden Markov model-derived structural alphabet for proteins: The learning of protein local shapes captures sequence specificity**. *Biochimica Et Biophysica Acta-General Subjects* 2005, **1724**(3):394-403.

122. Kolodny R, Koehl P, Guibas L, Levitt M: **Small libraries of protein fragments model native protein structures accurately**. *Journal of Molecular Biology* 2002, **323**(2):297-307.

123. Park BH, Levitt M: **The Complexity And Accuracy Of Discrete State Models Of Protein-Structure**. *Journal of Molecular Biology* 1995, **249**(2):493-507.

124. Ramachandran GN, Ramakrishnan C, Sasisekharan V: **STEREOCHEMISTRY OF POLYPEPTIDE CHAIN CONFIGURATIONS**. *Journal of Molecular Biology* 1963, **7**(1):95-&.

125. Klepeis JL, Androulakis IP, Ierapetritou MG, Floudas CA: **Predicting solvated peptide conformations via global minimization of energetic atom-to-atom interactions**. *Computers & Chemical Engineering* 1998, **22**(6):765-788.

126. Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA: **Energy Parameters In Polypeptides .10. Improved Geometrical Parameters And Nonbonded Interactions For Use In The Ecepp/3 Algorithm, With Application To Proline-Containing Peptides**. *Journal of Physical Chemistry* 1992, **96**(15):6472-6484.

127. Wetter M, Wright J: **Comparison Of A Generalized Pattern Search And A Genetic Algorithm Optimization Method**. In: *Eighth International IBPSA Conference.* Eindhoven, Netherlands; 2003.

128. Rhee YM, Pande VS: **Multiplexed-replica exchange molecular dynamics method for protein folding simulation**. *Biophysical Journal* 2003, **84**(2):775-786.

129. **Peptide Side Chain Conformational Analysis** [http://dunbrack.fccc.edu/bbdep/confanalysis.php]

130. Berg JM, Tymoczko JL, Stryer L: **Biochemistry, Fifth Edition**. New York: W. H. Freeman & Co.; 2002.

131. Case DA, Darden TA, T.E. Cheatham I, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M *et al*: **AMBER 9**. *University of California, San Francisco* 2006.

132. Ryckaert JP, Ciccotti G, Berendsen HJC: **Numerical Integration Of Cartesian Equations Of Motion Of A System With Constraints - Molecular-Dynamics Of N-Alkanes**. *Journal of Computational Physics* 1977, **23**(3):327-341.

133. Andrea TA, Swope WC, Andersen HC: **The Role Of Long Ranged Forces In Determining The Structure And Properties Of Liquid Water**. *Journal of Chemical Physics* 1983, **79**(9):4576-4584.

134. Delano WL: **PyMOL Molecular Graphics System (DeLano Scientific, Palo Alto, California, USA)**. In.; 2008.

135. Stanfield RL, Wilson IA: **Protein-Peptide Interactions**. *Current Opinion in Structural Biology* 1995, **5**(1):103-113.

136. Kauffman SA: **Origins of Order**. New York: Oxford University Press; 1993.

137. Vaughan TJ, Williams AJ, Pritchard K, Osbourn JK, Pope AR, Earnshaw JC, McCafferty J, Hodits RA, Wilton J, Johnson KS: **Human antibodies with sub-nanomolar affinities isolated from a large non-immunized phage display library**. *Nature Biotechnology* 1996, **14**(3):309-314.

138. Lipschultz CA, Li YL, Smith-Gill S: **Experimental design for analysis of complex kinetics using surface plasmon resonance**. *Methods* 2000, **20**(3):310-318.

139. Ekins S, Honeycutt JD, Metz JT: **Evolving molecules using multi-objective optimization: applying to ADME/Tox**. *Drug Discovery Today* 2010, **15**(11-12):451-460.

140. Arnaut L, Formosinho S, Burrows H: **Chemical Kinetics From Molecular Structure to Chemical Reactivity**. New York: Elsevier; 2007.

141. Fong CC, Wong MS, Fong WF, Yang MS: **Effect of hydrogel matrix on binding kinetics of protein-protein interactions on sensor surface**. *Analytica Chimica Acta* 2002, **456**(2):201-208.

142. Boehr DD, Wright PE: **How do proteins interact?** *Science* 2008, **320**(5882):1429-1430.

143. Lange OF, Lakomek NA, Fares C, Schroder GF, Walter KFA, Becker S, Meiler J, Grubmuller H, Griesinger C, de Groot BL: **Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution**. *Science* 2008, **320**(5882):1471-1475.

144. Tsai CJ, Ma BY, Sham YY, Kumar S, Nussinov R: **Structured disorder and conformational selection**. *Proteins-Structure Function and Genetics* 2001, **44**(4):418-427.

145. Baggio R, Carven GJ, Chiulli A, Palmer M, Stern LJ, Arenas JE: **Induced fit of an epitope peptide to a monoclonal antibody probed with a novel parallel surface plasmon resonance assay**. *Journal of Biological Chemistry* 2005, **280**(6):4188-4194.

146. Koshland DE: **The key-lock theory and the induced fit theory**. *Angewandte Chemie-International Edition* 1994, **33**(23-24):2375-2378.

147. Rini JM, Schulzegahmen U, Wilson IA: **Structural evidence for induced fit as a mechanism for antibody-antigen recognition**. *Science* 1992, **255**(5047):959-965.

148. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK: **Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners**. *Bmc Genomics* 2007, **9**.

149. Ito W, Yasui H, Kurosawa Y: **Mutations In The Complementarity-Determining Regions Do Not Cause Differences In Free-Energy During The Process Of Formation Of The Activated Complex Between An Antibody And The Corresponding Protein Antigen**. *Journal of Molecular Biology* 1995, **248**(4):729-732.

150. Ross PD, Subramanian S: **Thermodynamics Of Protein Association Reactions - Forces Contributing To Stability**. *Biochemistry* 1981, **20**(11):3096-3102.

151. Schreiber G, Fersht AR: **Interaction Of Barnase With Its Polypeptide Inhibitor Barstar Studied By Protein Engineering**. *Biochemistry* 1993, **32**(19):5145-5150.

152. Schreiber G, Fersht AR: **Rapid, electrostatically assisted association of proteins**. *Nature Structural Biology* 1996, **3**(5):427-431.

153. Wallis R, Leung KY, Pommer AJ, Videler H, Moore GR, James R, Kleanthous C: **Protein-Protein Interactions In Colicin E9 Dnase-Immunity Protein Complexes .2. Cognate And Noncognate Interactions That Span The Millimolar To Femtomolar Affinity Range**. *Biochemistry* 1995, **34**(42):13751-13759.

154. Kourentzi K, Srinivasan M, Smith-Gill SJ, Willson RC: **Conformational flexibility and kinetic complexity in antibody-antigen interaction**. *Journal of Molecular Recognition* 2008, **21**(2):114-121.

155. James LC, Roversi P, Tawfik DS: **Antibody multispecificity mediated by conformational diversity**. *Science* 2003, **299**(5611):1362-1367.

156. Grunberg R, Leckner J, Nilges M: **Complementarity of structure ensembles in protein-protein binding**. *Structure* 2004, **12**:2125-2136.

157. Berezhkovskiy LM: **On the kinetics of peptide binding to MHC proteins**. *Biophysical Chemistry* 1998, **71**(1):1-8.

158. Witt SN, McConnell HM: **The Kinetics Of Peptide Reactions With Class-Ii Major Histocompatibility Complex Membrane-Proteins**. *Accounts of Chemical Research* 1993, **26**(8):442-448.

159. Ferrante A, Gorski J: **Cooperativity of hydrophobic anchor interactions: Evidence for epitope selection by MHC class II as a folding process**. *Journal of Immunology* 2007, **178**(11):7181-7189.

160. Yaneva R, Schneeweiss C, Zacharias M, Springer S: **Peptide binding to MHC class I and II proteins: New avenues from new methods**. *Molecular Immunology* 2010, **47**(4):649-657.

161. Painter CA, Cruz A, Lopez GE, Stern LJ, Zavala-Ruiz Z: **Model for the Peptide-Free Conformation of Class II MHC Proteins**. *PLoS ONE* 2008, **3**(6).

162. Goldberg JM, Baldwin RL: **Kinetic mechanism of a partial folding reaction. 2. Nature of the transition state**. *Biochemistry* 1998, **37**(8):2556-2563.

163. Takeda S, McKay DB: **Kinetics of peptide binding to the bovine 70 kDa heat shock cognate protein, a molecular chaperone**. *Biochemistry* 1996, **35**(14):4636-4644.

164. Schmid D, Baici A, Gehring H, Christen P: **Kinetics of Molecular Chaperone Action**. *Science* 1994, **263**(5149):971-973.

165. Hoyer W, Hard T: **Interaction of Alzheimer's A beta peptide with an engineered binding protein - Thermodynamics and kinetics of coupled folding-binding**. *Journal of Molecular Biology* 2008, **378**(2):398-411.

166. Poulsen TR, Meijer PJ, Jensen A, Nielsen LS, Andersen PS: **Kinetic, affinity, and diversity limits of human polyclonal antibody responses against tetanus toxoid**. *Journal of Immunology* 2007, **179**(6):3841-3850.

167. Cohn M: **Degeneracy, mimicry and crossreactivity in immune recognition**. *Molecular Immunology* 2005, **42**:651-655.

168. Babaoglu K, Shoichet BK: **Deconstructing fragment-based inhibitor discovery**. *Nature Chemical Biology* 2006, **2**(12):720-723.

169. Hajduk PJ: **Puzzling through fragment-based drug design**. *Nature Chemical Biology* 2006, **2**(12):658-659.

170. Huang N, Jacobson MP: **Binding-Site Assessment by Virtual Fragment Screening**. *PLoS ONE* 2010, **5**(4).

171. Joseph-McCarthy D, Baber JC, Feyfant E, Thompson DC, Humblet C: **Lead optimization via high-throughput molecular docking**. *Current Opinion in Drug Discovery & Development* 2007, **10**(3):264-274.

172. Kawatkar S, Wang HM, Czerminski R, Joseph-McCarthy D: **Virtual fragment screening: an exploration of various docking and scoring protocols for fragments using Glide**. *Journal of Computer-Aided Molecular Design* 2009, **23**(8):527-539.

173. Zoete V, Grosdidier A, Michielin O: **Docking, virtual high throughput screening and in silico fragment-based drug design**. *Journal of Cellular and Molecular Medicine* 2009, **13**(2):238-248.

174. Hajduk PJ, Greer J: **A decade of fragment-based drug design: strategic advances and lessons learned**. *Nature Reviews Drug Discovery* 2007, **6**(3):211-219.

175. Tran HT, Pappu RV: **Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions**. *Biophysical Journal* 2006, **91**(5):1868-1886.

176. Legutki B, Magee DM, Stafford P, Johnston SA: **A general method for characterization of humoral immunity induced by a vaccine or infection**. *Vaccine* 2010, **28**:4529-4537.

177. Tapia V, Bongartz J, Schutkowski M, Bruni N, Weiser A, Ay B, Volkmer R, Or-Guil M: **Affinity profiling using the peptide microarray technology: A case study**. *Analytical Biochemistry* 2007, **363**(1):108-118.

178. Connors KA: **Binding Constants: The Measurement of Molecular Complex Stability**. New York: John WIley & Sons; 1987.

179. Mansfield SL, Jayawickrama DA, Timmons JS, Larive CK: **Measurement of peptide aggregation with pulsed-field gradient nuclear magnetic resonance spectroscopy**. *Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology* 1998, **1382**(2):257-265.

180. Bodmer JL, Schneider P, Tschopp J: **The molecular architecture of the TNF superfamily**. *Trends in Biochemical Sciences* 2002, **27**(1):19-26.

181. Eck MJ, Sprang SR: **The Structure Of Tumor Necrosis Factor-Alpha At 2.6-A Resolution - Implications For Receptor-Binding**. *Journal of Biological Chemistry* 1989, **264**(29):17595-17605.

182. Brune D, Kim S: **Predicting protein diffusion coefficients**. *PNAS* 1993, **90**:3835-3839.

183. Patron F, Adelman SA: **Solvent cage dynamics and chemical dynamics in liquids**. *Chemical Physics* 1991, **152**(1-2):121-131.

184. Rabinowitch E, Wood WC: **The collison mechanism and the primary photochemical process in solutions**. *Transactions of the Faraday Society* 1936, **32**(2):1381-1387.

185. London N, Movshovitz-Attias D, Schueler-Furman O: **The Structural Basis of Peptide-Protein Binding Strategies**. *Structure* 2010, **18**(2):188-199.

186. Vanhee P, Reumers J, Stricher F, Baeten L, Serrano L, Schymkowitz J, Rousseau F: **PepX: a structural database of non-redundant protein–peptide complexes**. *Nucleic Acids Research* 2009, **38**(Database Issue):D545-D551.

187. Vanhee P, Stricher F, Baeten L, Verschueren E, Lenaerts T, Serrano L, Rousseau F, Schymkowitz J: **Protein-Peptide Interactions Adopt the Same Structural Motifs as Monomeric Protein Folds**. *Structure* 2009, **17**(8):1128-1136.

188. Xu D, Tsai CJ, Nussinov R: **Hydrogen bonds and salt bridges across protein-protein interfaces**. *Protein Engineering* 1997, **10**(9):999-1012.

189. Nye TMW, Berzuini C, Gilks WR, Babu MM, Teichmann SA: **Statistical analysis of domains in interacting protein pairs**. *Bioinformatics* 2005, **21**(7):993-1001.

190. Henrick K, Thornton JM: **PQS: a protein quaternary structure file server**. *Trends in Biochemical Sciences* 1998, **23**(9):358-361.

191. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **A dissection of specific and non-specific protein - Protein interfaces**. *Journal of Molecular Biology* 2004, **336**(4):943-955.

192. Jones S, Marin A, Thornton JM: **Protein domain interfaces: characterization and comparison with oligomeric protein interfaces**. *Protein Engineering* 2000, **13**(2):77-82.

193. Lo Conte L, Chothia C, Janin J: **The atomic structure of protein-protein recognition sites**. *Journal of Molecular Biology* 1999, **285**(5):2177-2198.

194. Nooren IMA, Thornton JM: **Structural characterisation and functional significance of transient protein-protein interactions**. *Journal of Molecular Biology* 2003, **325**(5):991-1018.

195. Reynolds C, Damerell D, Jones S: **ProtorP: a protein-protein interaction analysis server**. *Bioinformatics* 2009, **25**(3):413-416.

196. Rodier F, Bahadur RP, Chakrabarti P, Janin J: **Hydration of protein-protein interfaces**. *Proteins-Structure Function and Bioinformatics* 2005, **60**(1):36-45.

197. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N: **Residue frequencies and pairing preferences at protein-protein interfaces**. *Proteins-Structure Function and Genetics* 2001, **43**(2):89-102.

198. Fong JH, Shoemaker BA, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV, Panchenko AR: **Intrinsic Disorder in Protein Interactions:**

**Insights From a Comprehensive Structural Analysis**. *PLOS Computational Biology* 2009, **5**(3).

199. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN: **Analysis of molecular recognition features (MoRFs)**. *Journal of Molecular Biology* 2006, **362**(5):1043-1059.

200. Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, Uversky VN, Dunker AK: **Characterization of molecular recognition features, MoRFs, and their binding partners**. *Journal of Proteome Research* 2007, **6**:2351-2366.

201. Babu MM: **NCI: a server to identify non-canonical interactions in protein structures**. *Nucleic Acids Research* 2003, **31**(13):3345-3348.

202. Beck DAC, Alonso DOV, Inoyama D, Daggett V: **The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(34):12259-12264.

203. Wang RX, Fang XL, Lu YP, Yang CY, Wang SM: **The PDBbind database: Methodologies and updates**. *Journal of Medicinal Chemistry* 2005, **48**(12):4111-4119.

204. Wang RX, Fang XL, Lu YP, Wang SM: **The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures**. *Journal of Medicinal Chemistry* 2004, **47**(12):2977-2980.

205. Huey R, Morris GM, Olson AJ, Goodsell DS: **A semiempirical free energy force field with charge-based desolvation**. *Journal of Computational Chemistry* 2007, **28**(6):1145-1152.

206. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ: **Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function**. *Journal of Computational Chemistry* 1998, **19**(14):1639-1662.

207. PDB: **Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description**. In.; 2008.

208. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M *et al*: **AMBER 9**: University of California, San Francisco; 2006.

209. **Atlas of Side-Chain and Main-Chain Hydrogen Bonding** [http://www.biochem.ucl.ac.uk/bsm/atlas/]

210. Dahiyat BI, Gordon DB, Mayo SL: **Automated design of the surface positions of protein helices**. *Protein Science* 1997, **6**(6):1333-1337.

211. **FIRST User Guide** [http://flexweb.asu.edu/software/first/user_guides/FIRST6.0_user_guide.pdf]

212. Gallivan JP, Dougherty DA: **Cation-pi interactions in structural biology**. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**(17):9459-9464.

213. Marshall MS, Steele RP, Thanthiriwatte KS, Sherrill CD: **Potential Energy Curves for Cation-pi Interactions: Off-Axis Configurations Are Also Attractive**. *Journal of Physical Chemistry A* 2009, **113**(48):13628-13632.

214. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R: **Optimal Docking Area: A New Method for Predicting Protein–Protein Interaction Sites**. *Proteins: Structure Function and Bioinformatics* 2005, **58**:134-143.

215. Protein Structural Analysis and Design Laboratory: **SLIDE — Screening for Ligands by Induced-fit Docking, User Guide**. In.; 2008.

216. Kuhn LA, Swanson CA, Pique ME, Tainer JA, Getzoff ED: **Atomic and residue hydrophilicity in the context of folded protein structures**. *Proteins-Structure Function and Genetics* 1995, **23**(4):536-547.

217. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA: **A second generation force field for the simulation of proteins, nucleic acids, and organic molecules**. *Journal of the American Chemical Society* 1995, **117**(19):5179-5197.

218. Mehler EL, Solmajer T: **Electrostatic Effects In Proteins - Comparison Of Dielectric And Charge Models**. *Protein Engineering* 1991, **4**(8):903-910.

219. Morris GM, Goodsell DS, Huey R, Olson AJ: **Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4**. *Journal of Computer-Aided Molecular Design* 1996, **10**(4):293-304.

220. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P: **A New Force-Field For Molecular Mechanical Simulation Of Nucleic-Acids And Proteins**. *Journal of the American Chemical Society* 1984, **106**(3):765-784.

221. Protein Data Bank: **Protein Data Bank Contents Guide v. 3.1**. In.; 2007.

222. **Amino Acid** [http://en.wikipedia.org/wiki/Amino_acid]

223. McDonald IK, Thornton JM: **Satisfying Hydrogen-Bonding Potential In Proteins**. *Journal of Molecular Biology* 1994, **238**(5):777-793.

224. Koide S, Sidhu SS: **The Importance of Being Tyrosine: Lessons in Molecular Recognition from Minimalist Synthetic Binding Proteins**. *Acs Chemical Biology* 2009, **4**(5):325-334.

225. Wang RX, Lu YP, Fang XL, Wang SM: **An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes**. *Journal of Chemical Information and Computer Sciences* 2004, **44**(6):2114-2125.

226. Bogan AA, Thorn KS: **Anatomy of hot spots in protein interfaces**. *Journal of Molecular Biology* 1998, **280**(1):1-9.

227. Fesik SW, Shuker SB, Hajduk PJ, Meadows RP: **SAR by NMR: An NMR-based approach for drug discovery**. *Protein Engineering* 1997, **10**:73-73.

228. Tsutsui Y, Wintrode PL: **Hydrogen/deuterium, exchange-mass spectrometry: A powerful tool for probing protein structure, dynamics and interactions**. *Current Medicinal Chemistry* 2007, **14**(22):2344-2358.

229. Schermann SM, Simmons DA, Konermann L: **Mass spectrometry-based approaches to protein-ligand interactions**. *Expert Review of Proteomics* 2005, **2**(4):475-485.

230. Pimenova T, Nazabal A, Roschitzki B, Seebacher J, Rinner O, Zenobi R: **Epitope mapping on bovine prion protein using chemical cross-**

**linking and mass spectrometry**. *Journal of Mass Spectrometry* 2008, **43**(2):185-195.

231. Schulz DM, Ihling C, Clore GM, Sinz A: **Mapping the topology and determination of a low-resolution three-dimensional structure of the calmodulin-melittin complex by chemical cross-linking and high-resolution FTICRMS: Direct demonstration of multiple binding modes**. *Biochemistry* 2004, **43**(16):4703-4715.

232. Sinz A, Kalkhof S, Ihling C: **Mapping protein interfaces by a trifunctional cross-linker combined with MALDI-TOF and ESI-FTICR mass spectrometry**. *Journal of the American Society for Mass Spectrometry* 2005, **16**(12):1921-1931.

233. Vaque M, Ardrevol A, Blade C, Salvado MJ, Blay M, Fernandez-Larrea J, Arola L, Pujadas G: **Protein-ligand docking: A review of recent advances and future perspectives**. *Current Pharmaceutical Analysis* 2008, **4**(1):1-19.

234. Hetenyi C, van der Spoel D: **Efficient docking of peptides to proteins without prior knowledge of the binding site**. *Protein Science* 2002, **11**(7):1729-1737.

235. Petsalaki E, Stark A, Garcıa-Urdiales E, Russell RB: **Accurate Prediction of Peptide Binding Sites on Protein Surfaces**. *PLOS Computational Biology* 2009, **5**(3):e1000335.

236. Namasivayam V, Gunther R: **PSO@AUTODOCK: A fast flexible molecular docking program based on swarm intelligence**. *Chemical Biology & Drug Design* 2007, **70**(6):475-484.

237. Namasivayam V, Gunther R: **Flexible Peptide-Protein Docking Employing PSO@AUTODOCK**. In: *Computational Biophysics to Systems Biology (CBSB08): 2008*: John von Neumann Institute for Computing; 2008: 337-340.

238. Corbeil CR, Moitessier N: **Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs**. *Journal of Chemical Information and Modeling* 2009, **49**(4):997-1009.

239. Englebienne P, Moitessier N: **Docking Ligands into Flexible and Solvated Macromolecules. 4. Are Popular Scoring Functions Accurate**

**for this Class of Proteins?** *Journal of Chemical Information and Modeling* 2009, **49**(6):1568-1580.

240. Ghersi D, Sanchez R: **Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites**. *Proteins-Structure Function and Bioinformatics* 2009, **74**(2):417-424.

241. Bordner AJ, Abagyan R: **Statistical analysis and prediction of protein-protein interfaces**. *Proteins-Structure Function and Bioinformatics* 2005, **60**(3):353-366.

242. Sanner MF, Olson AJ, Spehner JC: **Reduced surface: An efficient way to compute molecular surfaces**. *Biopolymers* 1996, **38**(3):305-320.

243. **Pymol Home Page** [http://pymol.sourceforge.net]

244. Wolfram Research Inc.: **Mathematica 7.0**. In. Champaign, IL; 2008.

245. **Pepsite** [http://pepsite.embl.de]

246. **ODA Server Submission** [http://www.molsoft.com/oda]

247. Lawlor MA, Alessi DR: **PKB/Akt: a key mediator of cell proliferation, survival and insulin responses?** *Journal of Cell Science* 2001, **114**(16):2903-2910.

248. Sale EM, Sale GJ: **Protein kinase B: signalling roles and therapeutic targeting**. *Cellular and Molecular Life Sciences* 2008, **65**(1):113-127.

249. **PI3K / AKT Signaling** [http://www.cellsignal.com/pathways/akt-signaling.jsp]

250. Goswami A, Ranganathan P, Rangnekar VM: **The phosphoinositide 3-kinase/Akt1/Par-4 axis: A cancer-selective therapeutic target**. *Cancer Research* 2006, **66**(6):2889-2892.

251. Osaki M, Oshimura M, Ito H: **PI3K-Akt pathway: Its functions and alterations in human cancer**. *Apoptosis* 2004, **9**(6):667-676.

252. Wang S, Basson MD: **Identification of functional domains in AKT responsible for distinct roles of AKT isoforms in pressure-stimulated cancer cell adhesion**. *Experimental Cell Research* 2007, **314**:286-296.

253. Lindsley CW, Barnett SF, Yaroschak M, Bilodeau MT, Layton ME: **Recent progress in the development of ATP-Competitive and**

**allosteric akt kinase inhibitors**. *Current Topics in Medicinal Chemistry* 2007, **7**(14):1349-1363.

254. Hiromura M, Okada F, Obata T, Auguin D, Shibata T, Roumestand C, Noguchi M: **Inhibition of Akt kinase activity by a peptide spanning the beta A strand of the proto-oncogene TCL1**. *Journal of Biological Chemistry* 2004, **279**(51):53407-53418.

255. Lippa B, Pan G, Corbett M, Li C, Kauffman GS, Pandit J, Robinson S, Wei L, Kozina E, Marr ES *et al*: **Synthesis and structure based optimization of novel Akt inhibitors**. *Bioorganic & Medicinal Chemistry Letters* 2008, **18**:3359-3363.

256. **Crosstide Peptide Substrate 1mM from Invitrogen** [http://www.biocompare.com/ProductDetails/381873/Crosstide-Peptide-Substrate-1mM.html]

257. Meier R, Alessi DR, Cron P, Andjelkovic M, Hemmings BA: **Mitogenic activation, phosphorylation, and nuclear translocation of protein kinase B beta**. *Journal of Biological Chemistry* 1997, **272**(48):30491-30497.

258. Freeman-Cook KD, Autry C, Borzillo G, Gordon D, Barbacci-Tobin E, Bernardo V, Briere D, Clark T, Corbett M, Jakubczak J *et al*: **Design of Selective, ATP-Competitive Inhibitors of Akt**. *Journal of Medicinal Chemistry* 2010, **53**(12):4615-4622.

259. Thomas CC, Deak M, Alessi DR, van Aalten DMF: **High-resolution structure of the pleckstrin homology domain of protein kinase B/Akt bound to phosphatidylinositol (3,4,5)-trisphosphate**. *Current Biology* 2002, **12**(14):1256-1262.

# APPENDIX A

# SUPPLEMENTAL MATERIALS

A publicly accessible repository for supplemental materials has been established at http://www.innovationsinmedicine.org/pprmint. The contents are listed below. "Readme" files are included in subdirectories where explanatory information is needed.

| **Subdirectory** | **File(s)** | **Description** |
|---|---|---|
| dataset | *.pdb | 3,924 curated, energy minimized PDB format files each containing a single peptide-protein interface, with peptide atoms designated as chain P and target atoms within 25Å of any P chain residue as chain I |
| database | pprmint.mdb | A Microsoft Access relational database containing results of analysis of the 3,924 PPRMint interfaces. See readme.pdf for details of tables and fields. |
| poptop | *.vb | Application source code for PopTop software (VB.net) |
| vspm | *.pse | Pymol session files corresponding to VSPM mapping of test set interfaces shown in figures, viewable using Pymol molecular viewer [134]. |
| vspm | vspm.zip | Zip archive containing application source code (9 Python programs) and readme.pdf describing usage, inputs, outputs, options, dependencies, and system requirements. |