OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Exploring Protein-Peptide Binding Specificity through Computational Peptide Screening

Arnab Bhattacherjee, Stefan Wallin*

Department of Astronomy and Theoretical Physics, Computational Biology and Biological Physics group, Lund University, Lund, Sweden

Abstract

The binding of short disordered peptide stretches to globular protein domains is important for a wide range of cellular processes, including signal transduction, protein transport, and immune response. The often promiscuous nature of these interactions and the conformational flexibility of the peptide chain, sometimes even when bound, make the binding specificity of this type of protein interaction a challenge to understand. Here we develop and test a Monte Carlo-based procedure for calculating protein-peptide binding thermodynamics for many sequences in a single run. The method explores both peptide sequence and conformational space simultaneously by simulating a joint probability distribution which, in particular, makes searching through peptide sequence space computationally efficient. To test our method, we apply it to 3 different peptide-binding protein domains and test its ability to capture the experimentally determined specificity profiles. Insight into the molecular underpinnings of the observed specificities is obtained by analyzing the peptide conformational ensembles of a large number of binding-competent sequences. We also explore the possibility of using our method to discover new peptide-binding pockets on protein structures.

Citation: Bhattacherjee A, Wallin S (2013) Exploring Protein-Peptide Binding Specificity through Computational Peptide Screening. PLoS Comput Biol 9(10): e1003277. doi:10.1371/journal.pcbi.1003277

Editor: Lilia M. Iakoucheva, University of California San Diego, United States of America

Received April 24, 2013; **Accepted** August 30, 2013; **Published** October 24, 2013

Funding: This work was supported by the Swedish Research Council/2007-6202 (http://www.vr.se/), the Royal Swedish Physiographic Society (http://www.fysiografen.se/), and the Swedish National Infrastructure for Computing. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: stefan@thep.lu.se

Introduction

Protein-peptide interactions are involved in wide range of cellular processes and are more common than originally thought. Disordered peptide segments, often found within longer regions of disorder in proteins, typically undergo a binding-induced folding transition upon contact with a target molecule such that a specific structure is assumed [1]. It is not uncommon, however, that significant conformational diversity persists even after binding [2–4]. Disordered regions in proteins play pivotal roles in controlling cellular signaling networks [5], protein subcellular localization [6,7], protein degradation [8], and post-translational modification [9,10]. Remarkably, a recent estimate suggests that as much as around 40% of all links in protein interaction networks are due to binding of short peptide segments of around 3–10 amino acids in length to protein domains [11].

An apparently general property of protein-peptide interactions is their promiscuous nature, i.e., certain peptide positions contribute very little (or not at all) to the binding affinity, and thus can accommodate various amino acid types, while other positions require specific amino acid types for binding [12–14]. Indeed, many domain families recognize sets of peptide sequences conforming to particular amino acid patterns, or linear motifs. For example, SH3 domains bind sequences containing P-X-X-P where X is any amino acid and P is proline [12], and PDZ domains target short sequence patterns occurring at the extreme C-terminal end of proteins [15]. More than 100 such different linear motifs are known [16], however, many remain to be discovered [11]. Putative new linear motifs can be found by mining for overrepresented sequence patterns in evolutionarily related proteins [17] or in unrelated proteins sharing a common functional characteristic [18–20]. These methods are, however, limited by weak statistical signals, and cannot discover peptide segments involved in very few interactions or those not conforming to linear motifs. Subtle variations in specificity among domain members beyond a simple motif are crucial to their biological function [21,22]. It is therefore of importance to understand the detailed molecular underpinnings of protein-peptide recognition. To this end, simulation methods at the atomic level have recently been employed, including different variants of docking [23–29], implicit- and explicit-water molecular dynamics [30–34], and Monte Carlo-based approaches [35–37].

Because of the promiscuous nature of protein-peptide interactions, determining peptide binding specificity profiles requires finding the binding free energy for a large number of different sequences. This can be computationally prohibitively expensive, especially since peptide chain entropy can contribute significantly to binding affinity [31]. In this work, we describe and test a theoretical framework for exploring, in an efficient and representative way, the combined sequence and conformational space of peptides interacting with a given peptide-binding pocket. In testing the method, we focus on 3 different PDZ domains with distinct peptide-binding specificity profiles. The method developed relies on the so-called multisequence Monte Carlo (MC) approach [38,39] in which a joint probability distribution in conformation and sequence space is simulated. Updates in conformation and sequence are performed as ordinary MC moves and thereby put on an equal footing. In particular, this makes search through

Author Summary

The interactions between proteins play a crucial role for almost every undertaking of a cell. Many of these interactions are mediated by the binding of relatively short unstructured polypeptide segments, or peptides, in one protein to well-folded domains in other proteins. Such protein-peptide interactions have some interesting and special properties, e.g., promiscuity, which means many different peptide sequences are able to bind the same protein domain. Peptides also often exhibit structural flexibility even after binding a protein. These special properties make it desirable, but also challenging, to simulate protein-peptide binding in atomistic detail for many different peptide sequences. To this end, we have developed a computational algorithm that simultaneously explores the structure of protein-peptide complexes and the amino acid sequences of the peptide. In particular, our algorithm allows binding-competent peptide sequences to be generated in direct relation to their binding strengths. We also explored the possibility of using our method to locate new peptide-binding pockets on protein structures. Computational algorithms such as the one developed here may pave the way to reveal the full complexity of protein-protein interaction networks used in cells.

sequence space fast compared to calculating binding free energies for peptide sequences one after another. In our scheme, a representative sample of strongly binding peptide sequences can be obtained because the conditional probability distribution of sequences given bound peptide conformations becomes biased according binding free energy weights, as schematically illustrated in Figure 1. A major advantage of our method is that the underlying equilibrium conformational ensembles are readily available, which can provide insight into the interplay between specificity and the peptide conformational dynamics. We also explore the possibility of employing our method to the discovery of peptide-binding pockets, given only a protein structure as input.

Methods

All-atom computational model

All calculations in this work are performed using the model in Ref. [36]. It is an implicit-solvent model combining an all-atom representation of the protein chain with an effective energy function taking into account the major contributions of protein interactions, hydrogen bonding, electrostatic attraction, and the hydrophobic effect [40]. The model was developed and tested based on the folding of small peptides and proteins and thereafter adapted particularly for protein-peptide binding [35,36]. The potential energy function can be decomposed into five terms,

$$E(\bar{r}) = E_{\rm ev} + E_{\rm loc} + E_{\rm hb} + E_{\rm sc} + E_{\rm des}\,, \qquad (1)$$

representing excluded-volume interactions, local backbone interactions, hydrogen bonding, sidechain-sidechain interactions, and a backbone desolvation effect, respectively [36]. Because of the effective nature of the energy function, assigning a physical unit to the energy E is not straightforward. We therefore use dimensionless units to express E and $k_{\rm B}T$, where T is the temperature. All bond lengths and angles are kept fixed at values derived from a statistical analysis of protein structures in the Protein Data Bank [41]. In addition, some torsional angles, such as the peptide bond angle $\omega = 180^\circ$, are kept fixed at "ideal" values. Therefore, the degrees of freedom of the model are a set of torsional angles and overall chain orientations. More precisely, a conformation of one or more protein chains, $\bar{r}$, is determined by the backbone torsional angles, $\phi_{\rm i}$, $\psi_{\rm i}$, a set of sidechain torsional angles, $\chi_{\rm i}$, for each amino acid i, and the overall rotational and translational orientation of each chain.

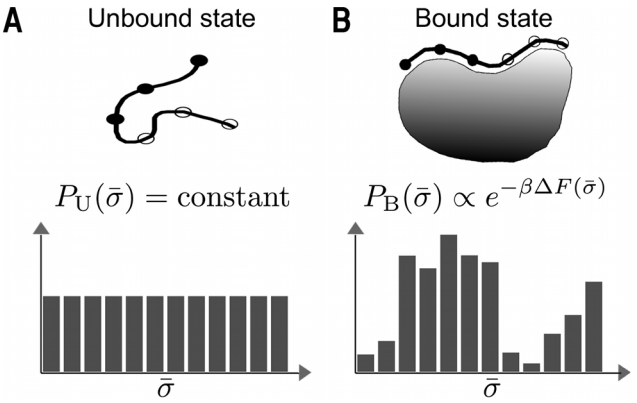

Figure 1. Schematic illustration of the computational peptide screening method. The method is based on simulating a joint probability distribution, $P(\bar{\sigma},\bar{r})$, where $\bar{\sigma}$ and $\bar{r}$ are amino acid sequence and chain conformation, respectively. Both conformational ($\bar{r}\rightarrow\bar{r}'$) and "mutational" ($\bar{\sigma}\rightarrow\bar{\sigma}'$) updates are performed as ordinary MC moves, subject to a Metropolis accept/reject question. Mutational updates are applied to a set of pre-defined variable amino acid positions on the peptide (open circles) while all other amino acids remain unchanged (filled circles). The procedure works in two steps. (A) In the first, iterative simulations of the unbound state (a free peptide) are performed creating a reference state where all $\bar{\sigma}$s occur with equal probability, i.e., the probability distribution $P_{\rm U}(\bar{\sigma})$ is flat. (B) In the second step, simulations of the protein-peptide bound state, B, are performed in which the distribution of $\bar{\sigma}$ becomes skewed according to the Boltzmann weights $e^{-\beta\Delta F(\bar{\sigma})}$, thereby favoring sequences with low binding free energies, $\Delta F(\bar{\sigma})$. The probability distribution $P_{\rm B}(\bar{\sigma})$ can be used to estimate relative $\Delta F(\bar{\sigma})$-values among the different sequences $\bar{\sigma}$ or give a representative view of the peptide-binding specificity of the protein.
doi:10.1371/journal.pcbi.1003277.g001

Protein-peptide binding specificity

Consider the interaction between a protein structure and a N-amino acid peptide with sequence $\bar{\sigma} = \{\sigma_1,...,\sigma_N\}$ where $\sigma_{\rm i}$ is the amino acid type of position i, and denote the chain conformation of the protein and peptide by $\bar{r}$. In principle, a complete description of the peptide-binding specificity of the protein means finding the binding free energy $\Delta F(\bar{\sigma}) = F_{\rm B}(\bar{\sigma}) - F_{\rm U}(\bar{\sigma})$, where $F_{\rm B}(\bar{\sigma})$ and $F_{\rm U}(\bar{\sigma})$ are the free energies of the bound (B) and unbound (U) states, respectively, for all possible $\bar{\sigma}$. This definition of binding free energy requires classifying conformations $\bar{r}$ as either B or U which can be done using some geometric criterion, such as the closeness of the peptide backbone to the peptide-binding pocket.

Binding free energy calculations are computationally intensive because they in principle require a full exploration of $\bar{r}$-space. For instance, the binding free energy of particular peptide at temperature T can be calculated using

$$\Delta F(\bar{\sigma}) = -\beta^{-1} \ln \frac{P(\mathrm{B}|\bar{\sigma})}{P(\mathrm{U}|\bar{\sigma})} \qquad (2)$$

where $P(\mathrm{B}|\bar{\sigma})$ and $P(\mathrm{U}|\bar{\sigma})$ are the probabilities of populating B and U, respectively, for a given peptide sequence $\bar{\sigma}$ at temperature

T, $\beta = 1/k_B T$, and k_B is Boltzmann's constant. Therefore, determining $\Delta F(\bar{\sigma})$ for many $\bar{\sigma}$ in a sequential manner is time consuming. As an alternative, we develop here a method that in a single run generates sequences from the probability distribution

$$P_B(\bar{\sigma}) = Z^{-1} \exp[-\beta \Delta F(\bar{\sigma})], \quad (3)$$

where Z is a normalization constant. Hence, rather than searching for a single optimally binding peptide, our method aims to "screen" for peptide sequences with low ΔF in a controlled manner thereby providing a representative picture of the peptide-binding specificity.

Multisequence Monte Carlo method for protein-peptide binding

The approach in this work for generating sequences according to the distribution in Equation 3 is based on the multisequence Monte Carlo method [38,39], meaning it relies on simulations of the joint probability distribution

$$P(\bar{r},\bar{\sigma}) = Z^{-1} e^{-\beta E(\bar{r},\bar{\sigma}) + g(\bar{\sigma})},$$
$$Z = \sum_{\bar{\sigma}} \int_{\bar{r}} d\bar{r} e^{-\beta E(\bar{r},\bar{\sigma}) + g(\bar{\sigma})}, \quad (4)$$

where $E(\bar{r},\bar{\sigma})$ is the potential energy of a conformation $\bar{r}$ with sequence $\bar{\sigma}$, and the sum and integral are taken over all $\bar{\sigma}$ and $\bar{r}$, respectively. Practically, this means designating a set of amino acid positions on the peptide as *variable*, for which the amino acid type is allowed to change dynamically through MC updates (see below for details). The parameters $g(\bar{\sigma})$ are important as they control the marginal distribution

$$P(\bar{\sigma}) = Z^{-1} \int_{\bar{r}} d\bar{r} e^{-\beta E(\bar{r},\bar{\sigma}) + g(\bar{\sigma})} = Z^{-1} Z(\bar{\sigma}) e^{g(\bar{\sigma})}, \quad (5)$$

where $Z(\bar{\sigma})$ is the canonical partition function for sequence $\bar{\sigma}$ at temperature T. We now make use of the division of $\bar{r}$-space into B and U states, such that $Z(\bar{\sigma}) = Z_B(\bar{\sigma}) + Z_U(\bar{\sigma}) = e^{-\beta F_B(\bar{\sigma})} + e^{-\beta F_U(\bar{\sigma})}$. This allows us to construct the probability distribution

$$P(s,\bar{\sigma}) = \begin{cases} Z^{-1} \int_{\bar{r}\in B} d\bar{r} e^{-\beta E(\bar{r},\bar{\sigma}) + g(\bar{\sigma})} = Z^{-1} e^{-\beta F_B(\bar{\sigma}) + g(\bar{\sigma})} & \text{if } s = B \\ Z^{-1} \int_{\bar{r}\in U} d\bar{r} e^{-\beta E(\bar{r},\bar{\sigma}) + g(\bar{\sigma})} = Z^{-1} e^{-\beta F_U(\bar{\sigma}) + g(\bar{\sigma})} & \text{if } s = U \end{cases}, \quad (6)$$

which can be used to construct a ratio,

$$\frac{P(B,\bar{\sigma})}{P(U,\bar{\sigma})} = e^{-\beta \Delta F(\bar{\sigma})}, \quad (7)$$

determined by $\Delta F(\bar{\sigma})$ and hence independent of the parameters $g(\bar{\sigma})$. Equation 7 shows that it is in principle possible to replace the sequential calculation of binding free energies $\Delta F(\bar{\sigma})$ for many $\bar{\sigma}$, by a *single* multisequence simulation of the distribution in Equation 4 and measuring the probabilities $P(B,\bar{\sigma})$ and $P(U,\bar{\sigma})$. Such an approach is possible but it also has practical limitations. The number of $P(B,\bar{\sigma})$ and $P(U,\bar{\sigma})$ quantities to be estimated grows exponentially (20^M) with the number of variable positions, M, meaning the approach is limited to very small M. The approach does in principle not depend on the parameters $g(\bar{\sigma})$ but in practice they would need to be carefully chosen to achieve sufficient sampling of sequence space, even for small M.

Computational peptide screening method

We do not pursue free energy calculations based directly on Equation 7 in this work. However, we take it as a starting point for developing our peptide screening method. First, we restate the probabilities in Equation 7 using Bayes' theorem,

$$P(B,\bar{\sigma}) = P(\bar{\sigma}|B)P(B),$$
$$P(U,\bar{\sigma}) = P(\bar{\sigma}|U)P(U), \quad (8)$$

where $P(B)$ and $P(U)$ are the total probabilities of occupying B and U (regardless of $\bar{\sigma}$), respectively, and $P(\bar{\sigma}|B)$ and $P(\bar{\sigma}|U)$ are conditional probabilities. Second, we make the choice $g(\bar{\sigma}) = \beta F_U(\bar{\sigma})$. This means that the distribution $P(U,\bar{\sigma})$, and hence $P(\bar{\sigma}|U)$, becomes flat (cf. Equations 6 and 8). We then obtain

$$P(\bar{\sigma}|B) \propto e^{-\beta \Delta F(\bar{\sigma})}, \quad (9)$$

and we can make the identification $P_B(\bar{\sigma}) \equiv P(\bar{\sigma}|B)$ (see Equation 3). To simplify our notation, we also put $P_U(\bar{\sigma}) \equiv P(\bar{\sigma}|U)$.

It is important to note that the conditional probabilities $P_B(\bar{\sigma})$ and $P_U(\bar{\sigma})$ are computationally convenient quantities because they do not depend on states $\bar{r}$ outside B and U, respectively. They can be obtained from separate multisequence simulations where $\bar{r}$ is restricted to B and U. We can now summarize our peptide screening method as a two-step strategy, illustrated in Figure 1:

1. *Unbound state simulation*. Determine $g(\bar{\sigma})$ parameter values such that all sequences $\bar{\sigma}$ occur with equal probability in a multisequence simulation of U, i.e., such that $P_U(\bar{\sigma})$ becomes flat.
2. *Bound state simulation*. Using the obtained $g(\bar{\sigma})$, perform a multisequence simulation of B. The generated sequences $\bar{\sigma}$ will become distributed according to the Boltzmann weights $e^{-\beta \Delta F(\bar{\sigma})}$.

Unbound state approximation

To further simplify our implementation of the above strategy we make the approximation that U consists of a free protein and a free peptide, without any interaction. There are then two contributions to the unbound state free energy $F_U(\bar{\sigma})$, a $\bar{\sigma}$-independent contribution from the protein, F_{free}^{prot}, and a $\bar{\sigma}$-dependent contribution from the free peptide, $F_{free}^{pep}(\bar{\sigma})$. We can ignore the quantity F_{free}^{prot}, putting $F_U(\bar{\sigma}) = F_{free}^{pep}(\bar{\sigma})$, because F_{free}^{prot} does not impact the distribution

$$P(\bar{\sigma}|U) \propto e^{-\beta F_{free}^{prot} - \beta F_{free}^{pep}(\bar{\sigma}) + g(\bar{\sigma})} \propto \int_{\bar{r}_{pep}\in peptide} d\bar{r}_{pep} e^{-\beta E(\bar{r}_{pep},\bar{\sigma}) + g(\bar{\sigma})}. \quad (10)$$

Hence, in calculations of the unbound state, we can rely on multisequence simulations of a free peptide chain ignoring the protein.

Linear model of the unbound state

A remaining question in implementing the strategy outlined above is how to determine the parameters $g(\bar{\sigma})$, such that they approximate well $\beta F_U(\bar{\sigma})$. We find that a simple linear form,

$$g(\bar{\sigma}) = h(\sigma_1) + \ldots + h(\sigma_N), \quad (11)$$

where $h(\sigma_i)$ depends on amino acid type, is sufficient to achieve a

good approximation. The 20 h-parameters can be interpreted as the contributions made to $\beta F_{\mathrm{U}}(\bar{\sigma})$ by the various amino acid types. This can be seen by considering the case in which the σ_{i}s are independent variables. In such a case, the unbound state free energy can be decomposed into position-independent contributions $f_{\mathrm{U}}(\sigma_{\mathrm{i}})$, i.e., $F_{\mathrm{U}}(\bar{\sigma}) = \sum_{\mathrm{i}=1}^{N} f_{\mathrm{U}}(\sigma_{\mathrm{i}})$, and the conditional probability distribution of $\bar{\sigma}$ given the unbound state can be written $P_{\mathrm{U}}(\bar{\sigma}) = \Pi_{\mathrm{i}=1}^{N} p_{\mathrm{U}}(\sigma_{\mathrm{i}})$, where

$$p_{\mathrm{U}}(\sigma_{\mathrm{i}}) \propto e^{-\beta f_{\mathrm{U}}(\sigma_{\mathrm{i}}) + h(\sigma_{\mathrm{i}})}. \tag{12}$$

Hence, the choice $g(\bar{\sigma}) = \beta F_{\mathrm{U}}(\bar{\sigma})$ amounts to setting $h(\sigma_{\mathrm{i}}) = \beta f_{\mathrm{U}}(\sigma_{\mathrm{i}})$. A good set of h-parameters can be found by iterative multisequence simulations of U (a free peptide) in which a flat distribution in sequence space, $p_{\mathrm{U}}(\sigma_{\mathrm{i}}) = \mathrm{constant}$, is eventually obtained. As seen from Equation 12, by measuring the probabilities $p_{\mathrm{U}}(\sigma_{\mathrm{i}})$ in simulations with an initial $h(\sigma_{\mathrm{i}})$ set, an improved set of values can be obtained by setting $h_{\mathrm{new}}(\sigma_{\mathrm{i}}) = h(\sigma_{\mathrm{i}}) - \ln p_{\mathrm{U}}(\sigma_{\mathrm{i}})$.

Using this procedure, we have determined $h(\sigma_{\mathrm{i}})$ parameters for 3 short peptides which provides approximately flat $P_{\mathrm{U}}(\bar{\sigma})$ distributions (see Figure S2 in Supporting Information). In particular, this shows that, despite the simplification, a linear approximation is sufficient to achieve $g(\bar{\sigma}) \approx \beta F_{\mathrm{U}}(\bar{\sigma})$ to a reasonably good approximation. This is important for the peptide screening method because it underlies the accuracy of Equation 9 which assumes $g(\bar{\sigma}) = \beta F_{\mathrm{U}}(\bar{\sigma})$. Errors in the approximation to the unbound state free energy will directly affect the conditional distribution, $P_{\mathrm{B}}(\bar{\sigma})$. More precisely, if $g(\bar{\sigma}) = \beta F_{\mathrm{U}}(\bar{\sigma}) + \delta(\bar{\sigma})$, then

$$P_{\mathrm{B}}(\bar{\sigma}) \propto e^{-\beta \Delta F(\bar{\sigma}) + \delta(\bar{\sigma})}. \tag{13}$$

The approximation errors $\delta(\bar{\sigma})$ are generally not possible to determine individually due to the size of the sequence space. An indication of the size of the errors can, however, be obtained from Figure S2. It shows that for the probability distributions of different amino acid types taken over all variable positions, the deviations are at most around 10%. Similar deviations from the desired Boltzmann distribution (see Figure 1B) for $P_{\mathrm{B}}(\bar{\sigma})$ should be expected. We also note that more elaborate approximations to $F_{\mathrm{U}}(\bar{\sigma})$ could easily be implemented, e.g., a position-dependent linear approximation with $20 \times M$ free parameters rather than the 20 parameters in Equation 11.

Errors introduced by the linear approximation on U would not impact free energy calculations performed using Equation 7, because this ratio is independent of the choice of $g(\bar{\sigma})$. Choosing $g(\bar{\sigma}) \approx \beta F_{\mathrm{U}}(\bar{\sigma})$ would nonetheless be a suitable choice for this method too, as a way to achieve good sequence space sampling.

Monte Carlo updates

The distribution in Equation 4 is realized through multisequence MC simulations. In these simulations, two different types of MC updates are included. Updates of the first type are conventional conformational updates ($\bar{r} \rightarrow \bar{r}'$) and include pivot moves, χ_{i}-angle rotamer turns, and rigid body rotation and translations, as described in previous work [35,36]. The second type of updates produces changes to the amino acid sequence of the peptide ($\bar{\sigma} \rightarrow \bar{\sigma}'$). These "mutational moves" are subject to an ordinary Metropolis accept/reject question, i.e., the new sequence is accepted with probability $P_{\mathrm{acc}} = \min[1, e^{-\Delta \mathcal{H}}]$, where $\Delta \mathcal{H} = \beta(E(\bar{r},\bar{\sigma}') - E(\bar{r},\bar{\sigma})) - g(\bar{\sigma}') + g(\bar{\sigma})$. Proposed sequences $\bar{\sigma}'$ are obtained by randomly picking a variable peptide position i and a new amino acid type, σ_{i}'. Thereafter, the peptide chain is rebuilt using the current $\bar{r}$, i.e., the set of ϕ_{i}-, ψ_{i}-, and χ_{i}-angles, and the new energy $E(\bar{r},\bar{\sigma}')$ calculated. A complication is that the number of actual degrees of freedom for different amino acid types differ. This can be handled by formally including two backbone angles, ϕ_{i} and ψ_{i}, and 5 side-chain angles, χ_{i}, as degrees of freedom for every variable amino acid position (7 is the maximum number of internal degrees of freedom for a residue in our model, occurring for lysine). For example, the geometry of an alanine residue is determined by two ϕ_{i}, ψ_{i} angles and a χ_{i} angle. This means that the potential energy $E(\bar{r})$ is independent of the remaining 4 χ_{i} angles, which will therefore quickly tend towards a uniform distribution. In a proposed mutation to an amino acid with additional (actual) degrees of freedom at position i, such as serine, the new amino acid will inherit the two ϕ_{i}, ψ_{i} angles and all 5 χ_{i} angles, which will determine its geometry. That detailed balance is indeed maintained by this scheme can be explicitly seen by comparing multisequence and a set of separate ordinary simulations of short peptides (see Figure S3 in Supporting Information).

Protein domains

The 3 peptide-binding proteins considered in this work are the 3rd PDZ domain of PSD-95, the 6th PDZ domain of GRIP1, and the PDZ domain of PICK1, which we refer to throughout the text as PSD95, GRIP1, and PICK1, respectively. Structures of peptide-bound complexes have been determined with X-ray crystallography for PSD95 (PDB id 1BE9) [42] and GRIP1 (1N7F) [43], and with NMR for PICK1 (2PKU) [44,45] (see Figure S1 in Supporting Information), with peptides sequences KQTSV, ATVRTYSC, and ESVKI, respectively.

Monte Carlo simulations

In order to test our peptide screening procedure (Figure 1), we perform also "fixed-sequence" simulations for comparison, following our earlier protocol [35,36]. This procedure explores the interaction between a given protein structure and a given peptide sequence in a straightforward way. The protein is kept close to an experimentally determined native structure through constraints on the C_{α}-atoms, leaving some backbone flexibility and complete sidechain flexibility. The peptide chain, by contrast, is left without constraints such that it can explore the entire protein surface. The protein and peptide chains are contained within a cubic box (side $L = 50$ Å) with periodic boundary conditions, corresponding to an effective concentration of ≈ 10 mM. To achieve an equilibrium picture of the interaction, the (dimensionless) simulation temperature is set such that both binding and unbinding events occur. In the present study, 10 independent fixed-sequence simulations of at least 7×10^8 MC steps were performed at $k_{\mathrm{B}}T = 0.45$ for each of the 9 PSD95-peptide pairs taken from Ref. [46].

Our peptide screening simulations (Figure 1B) differ from these fixed-sequence simulations in two ways. First, the peptide chain is restricted to the peptide-binding pocket of the protein using a constraint on the C_{α}-atom of the peptide C-terminal residue. This constraint is loose enough to still allow binding and "unbinding" of the peptide such that conformations in the bound state can be fully explored (for details, see section Peptide-binding pocket constraint). Second, in addition to the conformational MC updates for the protein and peptide chains, mutational updates are applied to the variable positions of the peptide (see above).

For PSD95, peptide screening simulations were performed with the peptides KKETE-x and KKE-xxx, where x indicates a variable amino acid position (derived from the sequence KKETEV which has been identified as a high affinity binder for PSD95 [46]). For GRIP1 and PICK1, simulations were performed

for ATVRT-*xxx* and ES-*xxx*, respectively. For each system with 3 variable positions we performed 20 independent runs, and for KKETE-*x* 3 independent runs were performed. All trajectories were at least 8×10^8 MC steps in length. The simulations were performed at $k_BT=0.45$, 0.55, and 0.51, for PSD95, GRIP1, and PICK1, respectively. These values were determined previously as midpoint temperatures for the different PDZ domains with their respective peptide ligands [36]. Simulations of the unbound state (step 1, Figure 1A) were performed for free peptide chains at the same respective temperatures. All multisequence simulations were initiated with $x=$alanine at the variable positions.

Peptide bound state

To monitor binding of the peptides in our simulations, we use a root-mean-square distance between the native and model peptide coordinates, $\mathbf{r}^{\mathrm{nat}}$ and $\mathbf{r}$, i.e.,

$$\mathrm{RMSD}=\sqrt{\frac{1}{N}\sum_{i}^{N}(r_i^{\mathrm{nat}}-r_i)^2}, \qquad (14)$$

where the sum goes over the N C_α-atoms of the peptide. The native peptide coordinates are taken from an experimentally determined structure (see Protein domains above). The peptide bound state (B) is defined as RMSD$<$6 Å, following Refs. [35,36].

Peptide-binding pocket constraint

To spatially constrain the peptide chain close to the peptide-binding pocket we use a simple constraint energy term, $E_{\mathrm{constr}}^{\mathrm{pep}}=k_{\mathrm{constr}}f(\delta_C)$, where $\delta_C=|\mathbf{r}_C^{\mathrm{nat}}-\mathbf{r}_C|$ is the deviation of the peptide C-terminal C_α-atom position, $\mathbf{r}_C$, from its position in the experimentally determined structure, $\mathbf{r}_C^{\mathrm{nat}}$. The function f is piecewise linear such that $f(x)=\max(0,x-10)$. This means that if the peptide C-terminal end moves more than 10 Å from its position in the native structure, there will be an energetic penalty. The constraint term was chosen in order to enhance sampling of the peptide bound state, without forbidding important bound state structures. The strength of the term is set to $k_{\mathrm{constr}}=10$.

Results/Discussion

In order to realize the computational peptide screening method imagined in Figure 1, a prerequisite is that relative binding free energies for different peptide sequences can be reasonably well estimated. We therefore start by testing our all-atom computational model for protein-peptide binding for predicting binding free energies on one of our 3 test domains. Second, we test the soundness of the developed screening method by comparing with the same binding free energy data. Third, we test the ability of the method to reproduce more generally the binding specificity profiles of the 3 domains and link them to conformational preferences of the peptide chain in the bound state. Lastly, we attempt "unrestricted" peptide screening in which the peptide is allowed to search freely the protein surface. Such an approach could potentially be used to locate new peptide-binding sites on protein structures.

All-atom computational model for protein-peptide binding

Previously, we have developed a MC-based approach for protein-peptide binding [35,36]. In this approach, the peptide is left free to explore the protein surface and relatively long simulations are performed such that a representative conformational ensemble can be obtained, including both bound and unbound states. The underlying all-atom model is taken from folding studies of proteins and was tested on a larger set of PDZ domains and peptides, particularly by comparing minimum-energy conformations with experimental structures of the protein-peptide complexes. In 8 out of 11 cases, the minimum-energy structures were within a root-mean-square distance RMSD (see Methods) of 6 Å from the experimental structures [35]. The method has also been used to study details of the peptide binding process by exploring the binding free energy landscapes for PDZ domains of different specificity classes [36].

We turn now to the ability of our model to quantitatively reproduce experimental binding affinity data. To this end, we use a study by Spaller *et al.* [46] in which isothermal titration calorimetry was used to determine binding affinities for the domain PSD95 and a number of peptide sequences under identical conditions. Of the peptides in Ref. [46], we focus on the 6-amino acid peptide KKETEV, a known high-affinity binder for PSD95, and 8 variants with modifications in either P_0 or P_{-2}; for PDZ peptide ligands, the C terminal position is denoted P_0 and the positions immediately upstream are P_{-1}, P_{-2}, P_{-3}, etc. Using our procedure [35,36], (see also Methods) we performed simulations of the interaction of each of these 9 peptides and PSD95. Binding free energies can be calculated in a straightforward way using Equation 2. To this end, we define the bound state, B, as peptide conformations with RMSD$<$6 Å, again following Refs. [35,36]. The experimentally measured dissociation constants, K_d, differ by approximately two orders of magnitude for the 9 peptides sequences, from 1.9 to 105 μM [46].

In Figure 2, experimental and calculated binding free energies are compared. There is a reasonably good agreement between the two sets of data, with the exception of one of the P_{-2} -variants which is predicted to bind too strongly. Excluding this outlier, the correlation is $r=0.86$. It should be noted that the comparison in Figure 2 involves relative binding free energies. Absolute binding free energies obtained from Equation 2 are generally different from those measured by Spaller *et al.* [46]. The reason is that our simulations are performed at a computationally convenient temperature where equilibrium can be reached, i.e., where both binding and unbinding are observed multiple times in a trajectory. The binding affinities obtained are relatively low ($K_d\sim$ mM) meaning that the simulation conditions used, $k_BT=0.45$, correspond to a higher temperature than the 298 K used in the experiments [46]. For the same reason, although the correlation between experimental and calculated binding free energies is good, the ranges observed are slightly different (approximately $4k_BT$ and $7k_BT$, respectively). The disagreement for the outlier sequence KKECEV, however, cannot be explained by temperature differences but rather indicates a limitation of our model in capturing all relevant energetics of the binding process. For class I PDZ domains, which includes PSD95, interactions at P_{-2} involve rather subtle intermolecular sidechain-sidechain hydrogen bonding (between the serine or threonine at P_{-2} and a histidine on the αB-helix of the PDZ domain [15]) that might not be captured entirely by our model. Overall, the result indicate that our model captures well variations in binding free energies among P_0 -variants while there are some limitations for P_{-2} -variants.

Testing the computational peptide screening method

We now turn to our computational peptide screening method. As illustrated by Figure 1, the method works by performing multisequence MC simulations in two steps. In these simulations,

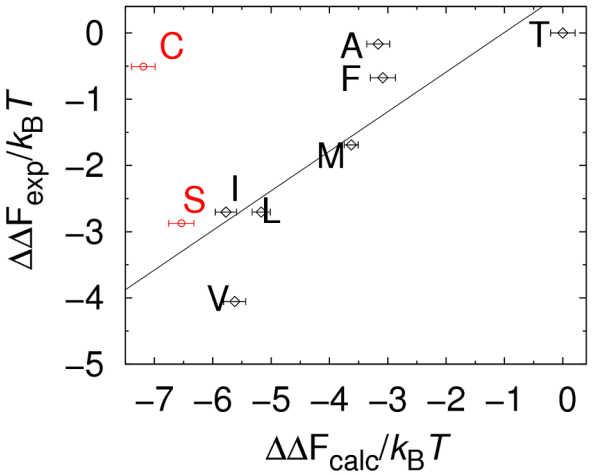

Figure 2. Comparing experimental and calculated relative binding free energies. As a quantitative test of the all-atom model [36] used in this work, we calculate binding free energies, ΔF_{calc}, for the protein PSD95 and 9 different peptide sequences. The ΔF_{calc} values were obtained from protein-peptide binding simulations performed separately for each PSD95-peptide pair according to our previous protocol [35,36] (see also Methods) and using Equation 2. The peptide sequences considered are derived from KKETEV (black diamond) and are either P_0 -variants, KKETE-[I/L/M/F/A/T], (black diamonds) or P_{-2} -variants, KKE[S/C]EV (red circles). All simulations were performed at $k_BT=0.45$ and standard errors were estimated from 10 independent runs. Experimental binding free energies, ΔF_{exp}, are taken from Ref. [46]. Both ΔF_{calc} and ΔF_{exp} values are shown relative to the weakest binding peptide. The solid line represents the best linear fit, exluding KKECEV, and the correlation coefficient is $r=0.86$.
doi:10.1371/journal.pcbi.1003277.g002

interlaced updates in conformational space and peptide sequence space (for variable peptide positions) are performed as ordinary MC updates. The first step involves iterative simulations of the unbound state (a free peptide chain) such that a reference state is created where all peptide sequences occur with roughly equal probabilities. In the second step, simulations of the peptide-bound state are performed where, by contrast, peptide sequences $\bar{\sigma}$ will be generated in a biased way according to the weights $e^{-\beta\Delta F(\bar{\sigma})}$. The theoretical background is described in detail in Methods.

As an initial test, we consider again PSD95 and screen the PDZ peptide-binding pocket using the peptide sequence KKETE-x, where x indicates a variable amino acid position. An example of a trajectory of the bound state B is shown in Figure 3B. Because the distribution of generated sequences is known ($\propto e^{-\beta\Delta F(\bar{\sigma})}$), the frequency of occurrence of different amino acid types at position x can be used to estimate the relative binding free energies for the 20 different sequences (i.e., KKETEG, KKETEA, KKETEV, etc.). These screening results can be directly compared with our results above, obtained for full protein-peptide simulations performed separately for different sequences. Our screening-derived binding free energies, $\Delta\Delta F_{screen}$, correlates well ($r=0.95$) with our previously obtained $\Delta\Delta F_{calc}$ values, as shown in Figure 3A. There are two approximations inherent to the peptide screening method. First, it is assumed that the unbound state consists of a free protein and a free peptide, without any interaction. Second, a linear approximation of the unbound state free energy is applied (see Methods for details). The agreement between the two different sets of results for PSD95 shown in Figure 3A means, in particular, that the approximations underlying the screening method do not strongly impact the results.

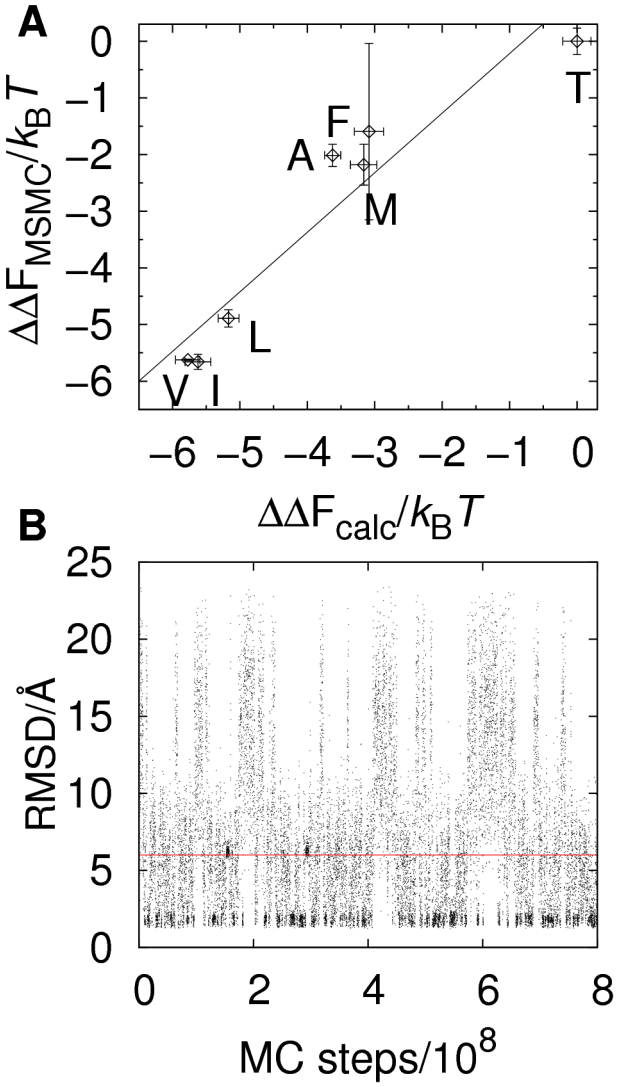

Figure 3. Relative binding free energies from computational peptide screening. We applied the peptide screening method to the protein PSD95 with the peptide KKETE-x, where x denotes a variable amino acid position. Relative binding free energies, $\Delta F_{screen}(\bar{\sigma})$, for the different possible peptide sequences $\bar{\sigma}$, were estimated from the distribution $P_B(\bar{\sigma})$ obtained in the second step of the screening procedure (see Figure 1) and using the relation $\Delta F_{screen}(\bar{\sigma})/k_BT \propto -\ln P_B(\bar{\sigma})$. (A) Comparison between ΔF_{calc} and ΔF_{screen} values for the P_0-variants in Figure 2. The correlation coefficient is $r=0.95$. Standard errors for ΔF_{screen} are estimated from 3 independent screening runs. As in Figure 2, the binding free energies are shown relative to the weakest binding peptide. (B) Example of a peptide screening run of PSD95 showing the evolution of RMSD, which measures the structural similarity of the peptide chain to the experimental structure, as a function of the number of elementary MC steps. The definition of bound state, RMSD<6 Å, is indicated by a horizontal line.
doi:10.1371/journal.pcbi.1003277.g003

Exploring peptide binding specificities

We now employ our screening method with the aim of more generally characterizing the peptide-binding specificity of a given protein domain. To this end, we apply the screening approach as in the previous section but now allow additional variable amino acid positions such that the main specificity-determining region of the peptide is covered. We focus on three different domains,

PSD95, GRIP1, and PICK1, each having a different specificity profile. PSD95 and GRIP1 are representative members of class I and II PDZ domains binding peptides with the sequence pattern $[T/S]-X-\Phi_{COOH}$ and $\Phi-X-\Phi_{COOH}$, respectively, where X is any residue, Φ is a hydrophobic residue, and COOH represents the peptide C terminus [15,47]. PICK1 is a domain with dual specificity, meaning that it binds peptides exhibiting either class I or II sequence patterns. We perform simulations in which 3 peptide positions are treated as variable, P_0, P_{-1}, and P_{-2}.

To identify possible binding motifs for a particular domain, we use weblogos [48] to represent the obtained conditional distribution of sequences, $P_B(\bar{\sigma})$. It is clear from Figure 4 that our result for GRIP1 is consistent with a class II domain, as expected from experimental data. The situations is less straightforward for PSD95. The position P_0 is occupied mainly by hydrophobic residues, particularly I, V, and L, and P_{-1} samples amino acid types almost uniformly. This in line with experimental results which have identified $[T/S]-X-[V/I/L]_{COOH}$ as the linear motif for PSD95 [46]. For P_{-2}, our simulations give only a weak signal for T/S which is likely due to the limitation of our model in capturing fully the binding energetics for this position for class I domains, as discussed above. It is interesting, however, that T7 phage display experiments [49] produced instances of hydrophobic residues at P_{-2}, particularly I and M, giving some support to our result in Figure 4A. This is also in line with the notion that the classification of PDZ domains is not strict and cross-interactions with other ligands are possible [22]. For PICK1, we find a specificity profile more closely related to class II rather than class I (see Figure 4C). This unexpected result suggests that PICK1 might be class II dominant despite its dual specificity nature. Further experiments will be needed to explore this possibility. It is at least partially supported by the study of Madsen *et al.* [50] where, using an assay based on fluorescence polarization, it was found that PICK1 showed a higher affinity for a class II than class I peptide. We also note that our screening method predominantly produce F residues at P_0 which is contrary to Ref. [50,51] where a preference for smaller hydrophobic residues was seen.

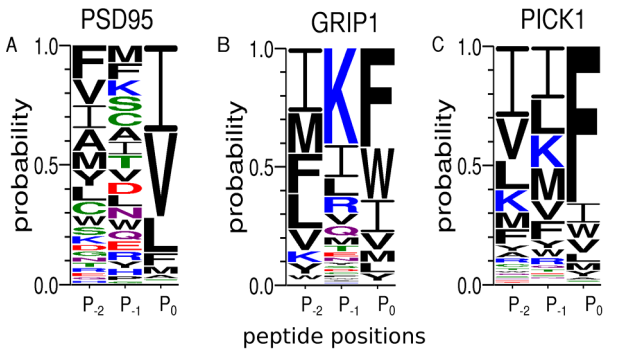

Figure 4. Peptide-binding specificity profiles from computational peptide screening. We applied our peptide screening method to the proteins (A) PSD95, (B) GRIP1, and (C) PICK1 with the peptides KKE-*xxx*, ATVRT-*xxx* and ES-*xxx*, respectively, where *x* indicates a variable amino acid position. As illustrated in Figure 1, the obtained probability distributions $P_B(\bar{\sigma})$ represent the peptide-binding specificities of the proteins. Shown are "one-dimensional" specificity profiles for the 3 variable peptide positions, illustrated as weblogos [48]. At a given peptide position i, the letter height is determined by $p_{B,i}(\sigma)$, i.e., the probability of observing the amino acid type σ at position i. The color scheme is as follows: hydrophobic (A, V, L, I, P, W, F, M, Y) black, polar (G, S, T, C) green, neutral (Q, N) purple, basic (K, R, H) blue, and acidic (D, E) red.
doi:10.1371/journal.pcbi.1003277.g004

The profiles in Figure 4 indicate that overall sequence variations are tolerated to different extents in the 3 different peptide positions. This is quantified in Figure 5, showing the sequence entropy, S_i, for the different amino acid positions. What is the reason for these differences? We note that S_i, which measures the degree of sequence randomness, is smallest for P_0 and largest for P_{-1} for all 3 domains. This is consistent with the general property of PDZ domains allowing only hydrophobic amino acids at P_0 while accommodating (mostly) any amino acid type at P_{-1}. Figure 5A shows that PSD95 exhibits a relatively high S_i also for P_{-2}. This is likely due to a limitation of our model to fully capture the T/S preference at this position, as discussed above. The slight preference for hydrophobic amino acids at P_{-2} is nonetheless interesting given the experimental support [49] for this observation. How can such atypical hydrophobic amino acids at P_{-2}, in some cases, be accommodated by a class I domain for which the recognized linear motif typically follows $[T/S]-X-\Phi_{COOH}$? In Figure 5D–F, we illustrate representative ensembles of bound peptide conformations for all three domains (regardless of peptide sequence, $\bar{\sigma}$) obtained from our screening simulations. PSD95 differ from the other two domains in that it displays a greater structural diversity of the peptide ligand. Such relatively major conformational flexibility is not uncommon for small ligands in complexes [52] and has been observed by our group previously for this PDZ domain [36]. In a somewhat simplified picture of PDZ-peptide binding, the bound state can be seen as a combination of two binding modes, where the peptide binds the domain either in a tight way, involving both P_0 and P_{-2}, or in a looser way, involving only P_0 (see Figure 5 and Figure S4 in the Supporting Information). This observation is in line with recently determined X-ray structures of class I PDZ domain-peptide complexes in which the peptides bind their respective domains mainly through P_0 and directed roughly perpendicular to the domain surface [53]. It is possible therefore that hydrophobic amino acids might be allowed at P_{-2}, in particular cases where the bound state include such "perpendicular" peptide conformations. Nonspecific hydrophobic contacts between P_{-2} sidechains and the domain surface might contribute to the stability of the complex. Such a picture would indeed explain the occurrence of some hydrophobic amino acids at P_{-2} both in our results and in phage-display experiments performed on PSD95 [49].

Discovery of peptide binding sites?

We have shown above that our peptide screening method can describe the gross features of the peptide binding specificities for a set of protein domains. The second step in our strategy (Figure 1B) involves multisequence simulations in which the peptide chain is artificially kept close to the peptide-binding pocket using a spatial constraint, in order to enhance the sampling of the bound state, B. Can the spatial constraint on the peptide be relaxed? The question is of interest because, if it turns out to be feasible, it opens up for using our screening method as a way to discover peptide-binding pockets on proteins, based on a 3-dimensional structure alone. To investigate the possibility for using our screening method for such structure-based binding-site discovery, we perform simulations of PSD95 following essentially the strategy in Figure 1, but with the difference that the peptide is left entirely unrestricted in the second step, i.e., it is free to diffuse in the simulation box and thus allowed to bind anywhere on the protein surface. Moreover, we make the 5 most C-terminal positions, P_0 to P_{-4}, variable. This approach is therefore truly unbiased in the sense that no prior knowledge is built in of either (1) the peptide-binding pocket on the protein or (2) which peptide sequences are binding competent.

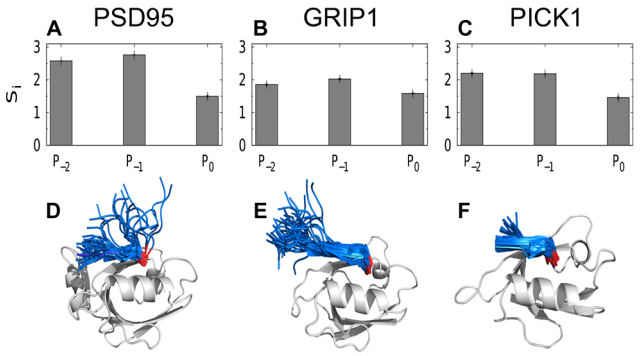

Figure 5. Interplay between peptide binding specificity and structural heterogeneity. A useful feature of the peptide screening method is that the underlying joint probability distribution $P(\bar{r},\bar{\sigma})$ (see Figure 1) can provide further insight into the structural underpinnings of specificity. Shown is a simple analysis of the specificity profiles of PSD95, GRIP1, and PICK1 in Figure 4. (A–C) The degree of sequence randomness at different peptide positions i, as measured by the sequence entropy $S_i = -\sum_\sigma p_{B,i}(\sigma)\ln p_{B,i}(\sigma)$, where $p_{B,i}(\sigma)$ is defined as in Figure 4 and the sum goes over all 20 different amino acid types σ. For reference, we note that for a position in which all amino acid types occur with equal probability, $S_i = \ln 20 \approx 3$. (D–F) Superposition of a random sample of bound state conformations $\bar{r}$ with various peptide sequences $\bar{\sigma}$, in ribbon representation (peptides shown in blue and PDZ domains in grey). For clarity, only single structures of the PDZ domains are shown. We find that the relatively larger structural heterogeneity at P_{-2} for peptides bound to PSD95 is connected to a higher S_i. The C-terminal amino acid, P_0, (red) is tightly bound to the peptide-binding pocket in all 3 cases and this feature is conserved across different binding-competent sequences.
doi:10.1371/journal.pcbi.1003277.g005

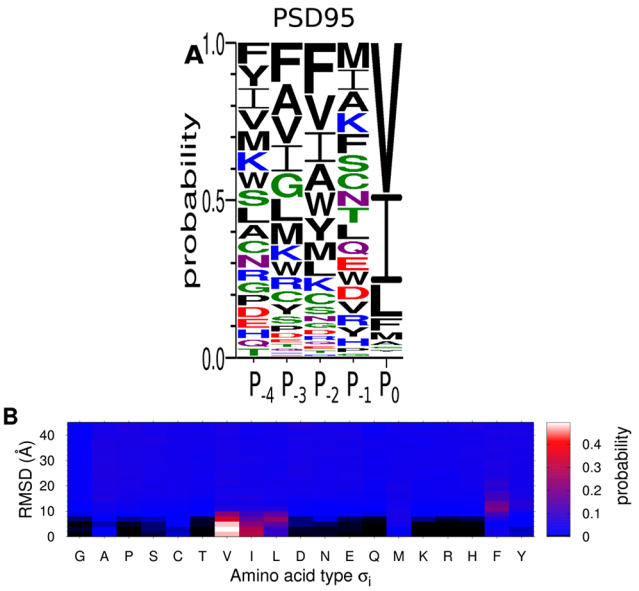

Figure 6. Using computational peptide screening for peptide-binding site discovery. We tested a modified version of the screening procedure in which the peptide is free to search the entire protein surface. This approach was applied to the protein PSD95 with the peptide G-*xxxxx* (*x* indicates a variable position), at $k_B T = 0.45$. (A) Peptide-binding specificity profile determined as previously for the 5 variable positions, i.e., using the conditional probability distribution of $\bar{\sigma}$ given the bound state, $P_B(\bar{\sigma})$. The letter color scheme is same as in Figure 4. (B) Probability distributions of $\bar{\sigma}$ given different values of RMSD, $P(\bar{\sigma}|\mathrm{RMSD})$, for the C-terminal peptide position, P_0. Far from the binding pocket (high RMSD), all amino acid types are visited roughly uniformly whereas close to the binding pocket (low RMSD) the distribution becomes skewed towards strongly binding sequences.
doi:10.1371/journal.pcbi.1003277.g006

Figure 6 shows the results of these unbiased peptide screening simulations. In Figure 6B, we show a probability distribution in terms of RMSD and amino acid type, σ_i, for P_0. The distribution should be interpreted such that, for a given RMSD, it gives the probability for the occurrence of various amino acid types at P_0. For example, at high RMSD values (>10 Å), the probability is roughly uniform ($\approx 1/20$), indicating that the peptide behaves as if in the unbound state. In particular, this means that peptide does not attach to surface regions other than the PDZ peptide-binding pocket despite the search through peptide sequence space. The picture changes drastically when the peptide is close to the peptide-binding pocket (low RMSD), in which the sequence distribution becomes skewed, particularly towards V, I, and L. In fact, the binding specificity profile for P_0, P_{-1}, and P_{-2}, constructed using the same bound state definition as before (RMSD<6 Å), is highly similar to the one obtained previously using the "restricted" screening simulations (cf. Figure 4A and Figure 6A). The consistency of these results suggests that our peptide screening method might be able to function as a tool to identify peptide binding sites on protein structures and at the same time provide a rough estimate of the their peptide-binding properties.

Summary and conclusions

We have developed an equilibrium MC-based method for characterizing protein-peptide interactions. The method samples jointly the peptide sequence and conformational spaces in a single run. In particular, this strategy makes search through sequence space computationally efficient and allows relative free energies to be estimated for a large number of peptides. In this work, we explored possible applications and used 3 different PDZ domains, with different peptide-binding specificities, as a test case. The peptide screening method relies on two approximations on the unbound state which are found not to impact the results significantly. Rather than measuring relative populations of bound and unbound states for many different peptide sequences, the method relies on measuring a conditional probability distribution of the sequences in a single run. Using this aspect of the method, we found good agreements with both full-scale protein-peptide binding simulations performed separately for each sequence as well as with experimental results. We also obtained specificity profiles for each of the 3 domains and compared with the experimentally known profiles, with a good overall agreement. An advantage of the method is that conformational ensembles are readily available for analysis, for visited sequences, which can reveal the interplay between binding specificity and conformational flexibility of the peptide chain. Finally, we explored the possibility of using the screening procedure for discovering new peptide-binding pockets on protein structures, with encouraging results.

Supporting Information

Figure S1 Experimental structures of the PSD95, GRIP1, and PICK1 domains in complex with peptide ligands. Visualization of the X-ray structures of (A) PSD95 [42] and (B) GRIP1 [43], and the NMR structure of (C) PICK1 [44]. The peptide ligands have the sequences KQTSV, ATVRTYSC, and ESVKI, respectively, and are shown in stick representation (deep blue, except the C-terminal amino acids shown in red). The

PDZ domains are shown in ribbon (light blue). The image was created using the PyMol molecular visualization program.
(TIFF)

Figure S2 Multisequence Monte Carlo simulations of free peptide chains. The first step of the peptide screening strategy (see Figure 1A) requires obtaining a uniform distribution of sequences, i.e., $P_U(\bar{\sigma}) = \text{constant}$. Multisequence simulations were performed of the isolated peptides KKE-*xxx* (PSD95), ATVRT-*xxx* (GRIP1), and ES-*xxx* (PICK1), where x represents a variable amino acid position, at $k_B T = 0.45$, 0.55, and 0.51, respectively. The figure shows probability distributions $p_U(\sigma)$ in amino acid type σ taken over all 3 variable positions. To achieve roughly flat distributions, i.e., $p_U(\sigma) \approx 1/20$, sets of 20 $h(\sigma)$ parameters were determined separately for each peptide by an iterative procedure, as explained in the text.
(TIFF)

Figure S3 Detailed balance in all-atom multisequence Monte Carlo simulations. To test the soundness of the proposed method, we performed multisequence simulations of the peptide A-*x*-A, where x is a variable amino acid position, at $k_B T = 0.45$. These simulations amounts to calculating the thermodynamic behavior for the 20 different variants x in a single run. For comparison, therefore, we perform ordinary MC simulations separately for each of the 20 tripeptides AGA, AAA, AVA, etc., at the same temperature. We test the consistency of the two set of results by comparing the probability distributions $p(\chi_i)$ for different sidechain rotamer angles, χ_i. The figure shows $p(\chi_i)$ for (A) χ_1, (B) χ_2, and (C) χ_3 for valine, obtained from the "fixed-sequence" simulation (FSMC) of the tripeptide AVA and the multisequence MC simulation (MSMC) of A-*x*-A, with $x = \text{V}$. The consistency of the results confirms that the multisequence simulation samples the correct thermodynamic distribution.
(TIFF)

Figure S4 Conformational diversity in the PSD95 peptide-bound state. Superposition of bound conformations from peptide screening simulations of PSD95, sub-grouped into peptides conformations with (A) RMSD <3 Å and (B) 3 Å $<$RMSD<6 Å, respectively.
(TIFF)

Author Contributions

Conceived and designed the experiments: AB SW. Performed the experiments: AB SW. Analyzed the data: AB SW. Contributed reagents/materials/analysis tools: AB SW. Wrote the paper: AB SW.

References

1. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6: 197–208.
2. Tompa P, Fuxreiter M (2008) Fuzzy complexes: polymorphism and structural disorder in proteinprotein interactions. Trends Biochem Sci 33: 2–8.
3. Borg M, Mittag T, Pawson T, Tyers M, Forman-Kay JD, et al. (2007) Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. Proc Natl Acad Sci USA 104: 9650–9655.
4. McDowell C, Chen J, Chen J (2013) Potential conformational heterogeneity of p53 bound to S100B($\beta\beta$). J Mol Biol 425: 999–1010.
5. Tantos A, Han KH, Tompa P (2012) Intrinsic disorder in cell signaling and gene transcription. Mol Cell Endocrinol 348: 457–465.
6. Subramani S (1992) Targeting of proteins into the peroxisomal matrix. J Membr Biol 125: 99–106.
7. Semenza JC, Hardwick KG, Dean N, Pelham HR (1990) ERD2, a yeast gene required for the receptor-mediated retrieval of luminal ER proteins from the secretory pathway. Cell 61: 1349–1357.
8. Young P, Deveraux Q, Beal RE, Pickart CM, Rechsteiner M (1998) Characterization of two polyubiquitin binding sites in the 26 S protease subunit 5a. J Biol Chem 273: 5461–5467.
9. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, et al. (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32: 1037–1049.
10. Gao J, Xu D (2012) Correlation between posttranslational modification and intrinsic disorder in protein. Pac Symp Biocomput 17: 94–103.
11. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, et al. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3: e405.
12. Li SS (2005) Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. Biochem J 390: 641–653.
13. Olsson N, Wallin S, James P, Borrebaeck CA, Wingren C (2012) Epitope-specificity of recombinant antibodies reveals promiscuous peptide-binding properties. Protein Sci 21: 1897–1910.
14. Eisen HN, Hou XH, Shen C, Wang K, Tanguturi VK, et al. (2012) Promiscuous binding of extracellular peptides to cell surface class I MHC protein. Proc Natl Acad Sci USA 109: 4580–4585.
15. Nourry C, Grant SG, Borg JP (2003) PDZ domain proteins: plug and play! Sci STKE 2003: RE7.
16. Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, et al. (2012) ELM-the database of eukaryotic linear motifs. Nucleic Acids Res 40: D242–251.
17. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, et al. (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202–208.
18. Rigoutsos I, Floratos A (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics 14: 55–67.
19. Neduva V, Russell RB (2006) DILIMOT: discovery of linear motifs in proteins. Nucleic Acids Res 34: W350–355.
20. Davey NE, Haslam NJ, Shields DC, Edwards RJ (2010) SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. Nucleic Acids Res 38: W534–539.
21. Zarrinpar A, Park SH, Lim WA (2003) Optimization of specificity in a cellular protein interaction network by negative selection. Nature 426: 676–680.
22. Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, et al. (2007) PDZ domain binding selectivity is optimized across the mouse proteome. Science 317: 364–369.
23. Desmet J, Wilson IA, Joniau M, De Maeyer M, Lasters I (1997) Computation of the binding of fully flexible peptides to proteins with flexible side chains. FASEB J 11: 164–172.
24. Liu Z, Dominy BN, Shakhnovich EI (2004) Structural mining: self-consistent design on flexible protein-peptide docking and transferable binding affinity potential. J Am Chem Soc 126: 8515–8528.
25. Tong JC, Tan TW, Ranganathan S (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex. Protein Sci 13: 2523–2532.
26. Niv MY, Weinstein H (2005) A flexible docking procedure for the exploration of peptide binding selectivity to known structures and homology models of PDZ domains. J Am Chem Soc 127: 14072–14079.
27. Petsalaki E, Stark A, Garcia-Urdiales E, Russell RB (2009) Accurate prediction of peptide binding sites on protein surfaces. PLoS Comput Biol 5: e1000335.
28. Raveh B, London N, Schueler-Furman O (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins 78: 2029–2040.
29. King CA, Bradley P (2010) Structure-based prediction of protein-peptide specificity in Rosetta. Proteins 78: 3437–3449.
30. Zacharias M, Springer S (2004) Conformational flexibility of the MHC class I alpha1-alpha2 domain in peptide bound and free states: a molecular dynamics simulation study. Biophys J 87: 2203–2214.
31. Basdevant N, Weinstein H, Ceruso M (2006) Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. J Am Chem Soc 128: 12766–12777.
32. Dhulesia A, Gsponer J, Vendruscolo M (2008) Mapping of two networks of residues that exhibit structural and dynamical changes upon binding in a PDZ domain protein. J Am Chem Soc 130: 8931–8939.
33. Ahmad M, Gu W, Helms V (2008) Mechanism of fast peptide recognition by SH3 domains. Angew Chem Int Ed Engl 47: 7626–7630.
34. Dagliyan O, Proctor EA, D'Auria KM, Ding F, Dokholyan NV (2011) Structural and dynamic determinants of protein-peptide recognition. Structure 19: 1837–1845.
35. Staneva I, Wallin S (2009) All-atom Monte Carlo approach to protein-peptide binding. J Mol Biol 393: 1118–1128.
36. Staneva I, Wallin S (2011) Binding free energy landscape of domain-peptide interactions. PLoS Comput Biol 7: e1002131.
37. Staneva I, Huang Y, Liu Z, Wallin S (2012) Binding of two intrinsically disordered peptides to a multi-specific protein: a combined Monte Carlo and molecular dynamics study. PLoS Comput Biol 8: e1002682.
38. Irbäck A, Peterson C, Potthast F, Sandelin E (1998) Monte Carlo procedure for protein design. Physical Review E 58: 5249–5252.

39. Irbäck A, Peterson C, Potthast F, Sandelin E (1999) Design of sequences with good folding properties in coarse-grained protein models. Structure 7: 347–360.
40. Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci USA 93: 13–20.
41. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, et al. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112: 535–542.
42. Doyle DA, Lee A, Lewis J, Kim E, Sheng M, et al. (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell 85: 1067–1076.
43. Im YJ, Park SH, Rho SH, Lee JH, Kang GB, et al. (2003) Crystal structure of GRIP1 PDZ6-peptide complex reveals the structural basis for class II PDZ target recognition and PDZ domain-mediated multimerization. J Biol Chem 278: 8501–8507.
44. Pan L, Wu H, Shen C, Shi Y, Jin W, et al. (2007) Clustering and synaptic targeting of PICK1 requires direct interaction between the PDZ domain and lipid membranes. EMBO J 26: 4576–4587.
45. Elkins JM, Papagrigoriou E, Berridge G, Yang X, Phillips C, et al. (2007) Structure of PICK1 and other PDZ domains obtained with the help of self-binding C-terminal extensions. Protein Sci 16: 683–694.
46. Saro D, Li T, Rupasinghe C, Paredes A, Caspers N, et al. (2007) A thermodynamic ligand binding study of the third PDZ domain (PDZ3) from the mammalian neuronal protein PSD-95. Biochemistry 46: 6340–6352.
47. Songyang Z, Fanning AS, Fu C, Xu J, Marfatia SM, et al. (1997) Recognition of unique carboxylterminal motifs by distinct PDZ domains. Science 275: 73–77.
48. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190.
49. Sharma SC, Memic A, Rupasinghe CN, Duc AC, Spaller MR (2009) T7 phage display as a method of peptide ligand discovery for PDZ domain proteins. Biopolymers 92: 183–193.
50. Madsen KL, Beuming T, Niv MY, Chang CW, Dev KK, et al. (2005) Molecular determinants for the complex binding specificity of the PDZ domain in PICK1. J Biol Chem 280: 20539–20548.
51. Bolia A, Gerek ZN, Keskin O, Banu Ozkan S, Dev KK (2012) The binding affinities of proteins interacting with the PDZ domain of PICK1. Proteins 80: 1393–1408.
52. Mobley DL, Dill KA (2009) Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". Structure 17: 489–498.
53. Elkins JM, Gileadi C, Shrestha L, Phillips C, Wang J, et al. (2010) Unusual binding interactions in PDZ domain crystal structures help explain binding mechanisms. Protein Sci 19: 731–741.