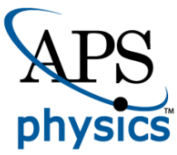

Peptide binding landscapes: Specificity and homophilicity across sequence space in a lattice model

Joohyun Jeon and M. Scott Shell

Phys. Rev. E **94**, 042405 — Published 10 October 2016

DOI: 10.1103/PhysRevE.94.042405

Peptide binding landscapes: specificity and homophilicity across sequence space in a lattice model

Joohyun Jeon and M. Scott Shell

Department of Chemical Engineering, University of California Santa Barbara, Santa Barbara, California 93106-5080

September 12, 2016

ABSTRACT

Peptide aggregation frequently involves sequences with strong *homophilic* binding character, i.e., sequences that self-assemble with like species in a crowded cellular environment, in the face of a multitude of other peptides or proteins as potential *heterophilic* binding partners. What kinds of sequences display a strong tendency towards homophilic binding and self-assembly, and what are the origins of this behavior? Here, we consider how sequence-specificity in oligomerization processes plays out in a simple 2D lattice statistical-thermodynamic peptide model that permits exhaustive examination of the entire sequence and configurational landscapes. We find that sequences with strong self-specificities have either alternating hydrophobic and hydrophilic residues or short patches of hydrophobic residues, both which minimize intramolecular hydrophobic interactions in part due to the constraints of the 2D lattice. We also find that these specificities are highly sensitive to entropic and free energetic features of the *unbound* conformational state, such that direct binding interaction energies alone do not capture the complete behavior. These results suggest that the ability of particular peptide sequences to self-assemble and aggregate in a many-protein environment reflects a precise balance of direct binding interactions and behavior in the unbound (monomeric) state.

I. INTRODUCTION

It is well-known that many short protein sequences can partially unfold and subsequently transition from soluble forms to amyloid fibrils [1], which are linked with a variety of serious human health disorders, including Alzheimer's disease, Parkinson's disease, and type II diabetes [2]. A particularly striking question is how these peptides manifest pronounced *self*-affinity, binding preferentially to their identical copies to form fibrils in a complex cellular environment with thousands of types of proteins in close vicinity [3]. For example, Krebs et. al. [4] showed that fibril growth is accelerated only when the seeding fibrils have a high degree of sequence similarity, and this is consistent with the fact that fibrils in each amyloid disease are typically composed of a single, dominant type of protein. Recently, this intrinsic sequence specificity was also used to create antibodies that are able to detect amyloid fibrils [5,6]. Specifically, antibodies grafted with known amyloidogenic sequences could successfully bind to small amyloidogenic oligomers and fibrils in vitro using their sequence-specific homophilic interactions.

While peptide aggregation propensities generally increase with hydrophobicity and indeed many predictive aggregation tools place a high weight on net sequence hydrophobicity [7], the mere presence of significant hydrophobic interactions does not explain the *homophilic* specificity of amyloid-forming peptides that induces their self-binding property. Indeed, there have been many efforts to predict or design amyloidogenic sequences [8–10], but a basic physical understanding of the emergence of homophilic specificity remains incomplete. Trovato et al. [11] used knowledge-based potentials to examine the most favorable binding register between two copies of the same (amyloidogenic) peptide, and discovered that in-register parallel alignment is nearly always preferred – which provides initial insight into the homophilicity problem. However, a much broader consideration of binding partners across sequence space can address many outstanding key questions, such as: which amino acid sequences promote high homophilic binding? What features of those sequences bear out homophilic specificity?

Can an examination of homophilicity contribute to a deeper understanding of well-known aggregation-prone sequences and sequence motifs?

Here, we examine the issue of homophilicity in a simple lattice protein model that offers a first approximation to the main driving forces present in polypeptides. Specifically, we consider the venerable HP two-dimensional lattice heteropolymer model, which involves only two types of residues, H (hydrophobic) and P (hydrophilic). A negative H-H contact energy mimics hydrophobic interactions, which are the main driving force for folding and for shaping the conformational ensemble [12]. For modest chain lengths, the model permits exploration of the full conformational and sequence spaces [13], while still capturing folding behavior and the possibility of a unique native conformation for some sequences [14]. Moreover, a binary sequence model contains some essential features of peptide self-assembly; for example, it was found experimentally that binary patterning of hydrophobic and hydrophilic residues contributed greatly to the design of aggregating peptide sequences [15–17]. The use of this particular model enables exact numerical calculation of thermodynamic quantities (e.g. free energies of folding and binding) and importantly, exhaustive examination of the entire sequence and binding spaces. The binding *landscape* of a given peptide then compares the binding equilibrium constant (or free energy) with itself to that with all other sequences, and thus bears out the relative degree of self-binding affinity and specificity. While certainly the model lacks many details that govern real protein self-assembly, it nonetheless gives rise to complex behavior from a set of well-defined and simple set of interaction rules – and thus provides a first understanding of how basic molecular driving forces translate to the binding landscape.

We explore the nature of the binding landscape in this model by characterizing the binding affinities between every possible pair of amino acid sequences of a given length. To do so, we compute free energies of both the bound and monomeric states using numerical partition function sums. We find that the binding affinity between two peptides is strongly influenced by the unfolding free energies (i.e., the free energy associated with each peptide adopting a linear binding pose) in addition to the direct

binding interactions. Further, we find that the unfolding free energy can be understood in terms of simple sequence patterns rather than sequence composition (i.e., fraction of hydrophobic residues). Sequences with the highest self-binding affinity that simultaneously maintain strong self-specificity are the alternating P/H sequence and those with hydrophobic residues that are clustered rather than distributed throughout the sequence. Both cases effect a low unfolding free energy cost, and are influenced by the geometric constraints of the 2D lattice.

II. METHODS

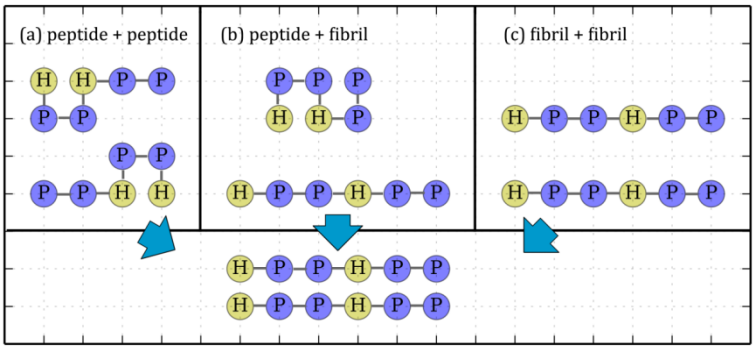

Figure 1. Illustration of peptide binding scenarios on a two dimensional square lattice: (a) peptide binding to peptide, (b) peptide binding to linear peptide (i.e., fibrillar peptide partner), and (c) linear peptide binding to linear peptide (i.e., binding of two fibrillar peptides). H and P circles represent hydrophobic and hydrophilic amino acid residues, respectively. In (a) and (b), peptides initially exist in an ensemble of partially folded states promoted by intramolecular H-H interactions (H-H contacts); upon binding in a linear arrangement, intermolecular H-H (H-H contacts) interactions promote affinity.

In the 2D lattice model, a peptide is a linear chain of N amino acids, and chain conformations are self-avoiding walks in a square lattice [13]. Interactions among residues are pairwise only between nearest neighbors. Energies are expressed in dimensionless units, with a favorable contact energy of -1 for H-H contacts; all other contacts (H-P and P-P) do not contribute energetics. We also include a sequence-independent hydrogen bonding interaction for intermolecular peptide association, described below.

Temperatures are expressed as dimensionless, such that k_B=1. When the peptide is in the monomeric state, it generally experiences a diverse set of conformations sampled according to Boltzmann populations, although some sequences will exhibit folding behavior and populate a particular conformation with high probability. Nonetheless, here we term the monomeric ensemble the "folded" state, even if it is a diverse conformational collection, because peptides must then "unfold" to a linear conformation to bind.

We model the binding of two peptides as a two-step process: the peptides first "unfold" to a linear binding pose, followed by co-alignment in either a parallel or anti-parallel fashion (Fig. 1 (a)). The binding affinity in this process is the free energy change from the folded state to the bound state, and therefore, is defined as the difference between the bound free energy and the sum of the two folding free energies of the two peptides. We restrict our study to linear and fully in-register binding cases. In reality, binding can occur between partially folded peptides or proteins [1], which occurs when the peptide length is large and the cost of unfolding is very high. We do not model such events because for very short sequences, we expect any partially folded binding pose to have significantly reduced affinity owing to the reduction in direct intermolecular interactions. We ignore off-register linear binding for the same reason.

We enumerate the full conformational space of the peptides and numerically compute the free energy of the monomeric or "folded" state, given by the partition sum,

$$A_{\text{fold}} = -T \ln \sum_i e^{-E_i/T} \tag{1}$$

where the index i proceeds over all conformations, E_i is the energy of conformation i, and T is the temperature. All terms in (1) and subsequent equations are dimensionless. For calibration, the average folding energy over all sequences approaches 1% of its minimum when T=0.1 and 1% of its maximum when T=10, as shown in Figure 2. The temperature at which the average reaches its midpoint value is T=0.53. We choose T=0.3 for our examination of binding, which gives a broad range of folding behavior over all of sequence space.

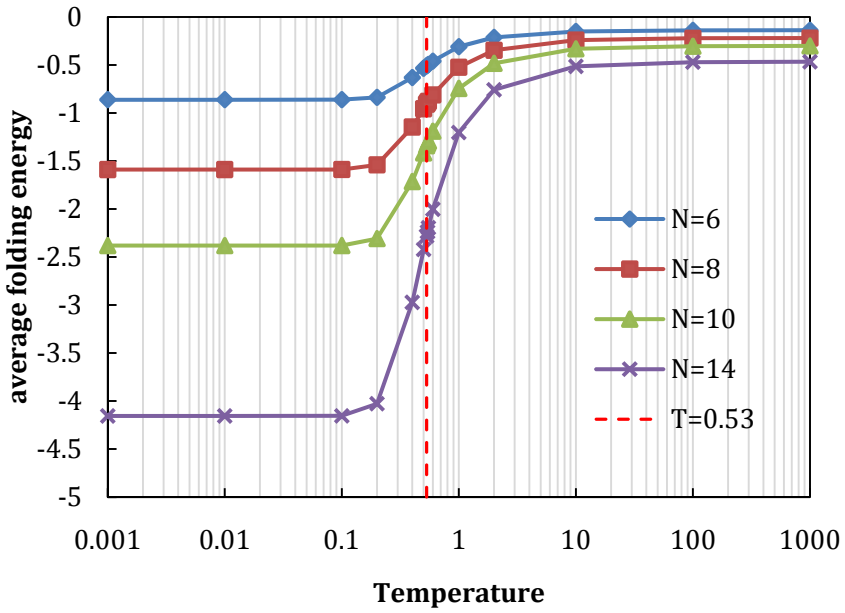

Figure 2. Average folding energy of all sequences with N=6, 8, 10, and 14 with respect to temperature. An effective folding temperature for the entire sequence ensemble, T_{fold}, is determined from the midpoint of the folding energy using $\langle E\rangle_{fold}(T_{fold}) = \left[\langle E\rangle_{fold}(T \to 0) + \langle E\rangle_{fold}(T \to \infty)\right]/2$.

The free energy of the bound state involves contributions from direct peptide-peptide hydrophobic interactions in both parallel and antiparallel poses. In addition to the H-H contact potential, we include a sequence-independent hydrogen bonding energy to capture differences in the number of hydrogen bonds that can be formed in the folded and bound states. Indeed, it has been suggested that while hydrophobic interactions remain a dominant driving force, hydrogen bonds may play more important roles in amyloid proteins than for soluble proteins [18]. We approximate hydrogen bonding energies with a term proportional to the peptide length N and temperature T. The expression for the bound free energy is

$$A_{\text{bound}} = -T \ln\left[e^{-(E_{\text{parallel}}+E_{HB})/T} + e^{-(E_{\text{antiparallel}}+E_{\text{HB}})/T}\right] \quad (2)$$

where E_{parallel} and $E_{\text{antiparallel}}$ give the hydrophobic contact energies in the parallel and antiparallel poses, and where E_{HB} is given by

$$E_{\mathrm{HB}} = -0.3\,N \tag{3}$$

Note that the hydrogen bonding term is trivially factored out of the logarithm such that A_{bound} is shifted directly by it.

With these formalities, the binding *affinity* of a sequence S to another sequence S' can be defined as the following free energy difference, per the path shown in Fig. 1:

$$\mathrm{Affinity}(S, S') = A_{\mathrm{bound}}(S, S') - A_{\mathrm{fold}}(S) - A_{\mathrm{fold}}(S') \tag{4}$$

Note that a lower (more negative) value of affinity indicates a stronger binding interaction. The self-binding affinity, which reflects the homophilic preference of a sequence S, follows as

$$\mathrm{Affinity}_{\mathrm{self}}(S) = A_{\mathrm{bound}}(S, S) - 2A_{\mathrm{fold}}(S) \tag{5}$$

We explore the entire sequence space, the set of all possible *N*-length sequences of H and P residues, which contains 2^N possible sequences in principle. However, we only consider symmetrically unique sequences because the N and C peptide termini are not distinguishable in the lattice model. Then, for example, the number of $(\mathrm{P/H})_{12}$ sequences is 2080 for N=12. Many natural peptide binders interact with their partners through a core composed of 5 to 15 residues [19] and therefore we examine lengths within this range. However, there are very few qualitative differences for our results towards the upper end of the range, and so we focus on N=12 as a representative case study.

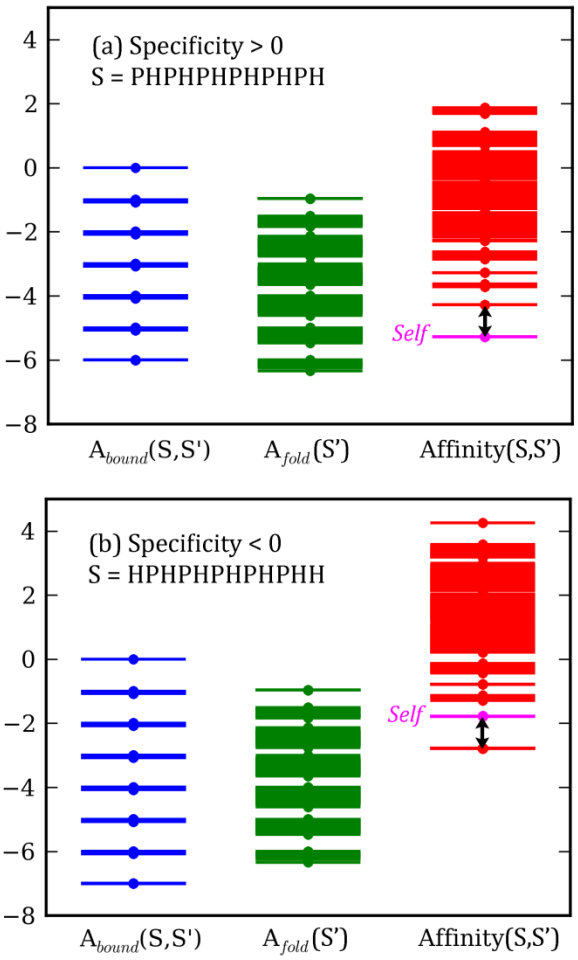

Figure 3. Illustrative energy level diagrams for the bound free energy, folded free energy, and binding affinity of two selected sequences that have (a) good and (b) poor specificity. Here we consider a low temperature (T=0.1) that makes the levels easier to distinguish visually. Each line represents a potential binding partner S' for target sequence S. In case (a), the specificity is positive and defined as the gap between the self-binding and next highest affinities. In case (b), the specific is negative and defined as the gap between the self-binding and lowest affinity.

We compute the folding free energies for all 2080 sequences for N=12, and the corresponding bound-state free energies and then affinities for all 2,162,160 sequence pairs. Figure 3 shows representative energy level diagrams for these quantities in which the "levels" span the full space of amino acid sequences for a potential binding partner S' interacting with target sequence S. For each target, one of the sequences S' has the lowest binding affinity to it. If this sequence is the same as S, then

the sequence is highly homophilic (Fig. 3 (a)). In other words, S binds well to itself and this specificity can be measured as the gap between the self-binding affinity and the next lowest binding affinity. If $S \neq S'$ then the case is non-homophilic, and the specificity is defined as the binding affinity gap between the self-binding affinity and the lowest binding affinity (Fig. 3 (b)). Therefore, the sign of the specificity indicates whether or not a peptide has high homophilic binding stability (positive if this is the case), and the degree of specificity is quantified by binding affinity gap [20,21]. Note that the affinity levels may be degenerate – more than one sequence may lie at each level. For sequences where the lowest affinity level is degenerate, we consider their specificity level to be zero.

To make these statements precise, the specificity is defined as

$$\begin{aligned}\text{Specificity}(S) &= \text{Affinity}\left(S, S'\right) - \text{Affinity}_{\text{self}}(S) \\ &= [A_{\text{bound}}(S, S') - A_{\text{bound}}(S, S)] - [A_{\text{fold}}(S') - A_{\text{fold}}(S)]\end{aligned} \tag{6}$$

where

$$S' = \begin{cases}\text{the sequence with the next lowest affinity} & \text{if } \text{Affinity}_{\text{self}} \text{ is the lowest} \\ \text{the sequence with the lowest affinity} & \text{otherwise}\end{cases} \tag{7}$$

This approach to characterizing specificity can be extended to the case of fibril-growth by monomeric addition, in which a peptide binds to a conformationally rigid fibril (Fig. 1 (b)). Indeed most peptides do not induce conformational changes in large partners upon binding, so as to minimize the entropic cost of binding [19]. In this case, the affinities can be written as

$$\text{Affinity}(S, S') = A_{\text{bound}}(S, S') - A_{\text{fold}}(S') \tag{8}$$

$$\text{Affinity}_{\text{self}}(S) = A_{\text{bound}}(S, S) - A_{\text{fold}}(S) \tag{9}$$

$$\text{Specificity} = [A_{\text{bound}}(S, S') - A_{\text{bound}}(S, S)] - [A_{\text{fold}}(S') - A_{\text{fold}}(S)] \tag{10}$$

Lastly, when two fibrils bind to each other or when no conformational changes happen upon binding (Fig. 1 (c)), the binding affinity is simply equal to the bound-state free energy because there is no unfolding cost, and specificity is then related to differences between bound-state free energies. The relations are

$$\text{Affinity}(S, S') = A_{\text{bound}}(S, S') \tag{11}$$

$$\text{Affinity}_{\text{self}}(S) = A_{\text{bound}}(S, S) \tag{12}$$

$$\text{Specificity} = A_{\text{bound}}(S, S') - A_{\text{bound}}(S, S) \tag{13}$$

III. RESULTS

A. What determines favorable peptide self-binding?

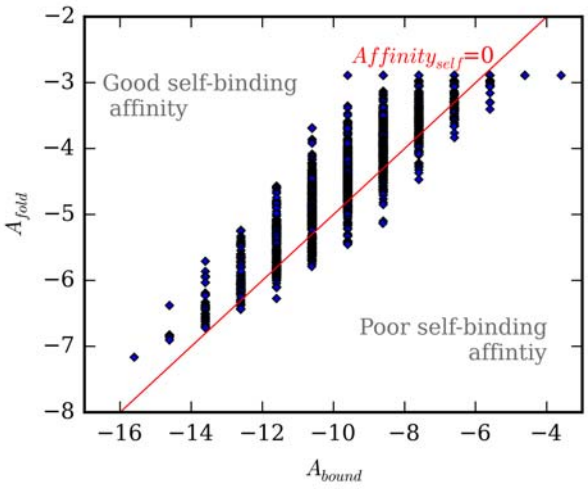

Figure 4. Self-binding and folding free energies of all 2080 $(P/H)_{12}$ sequences. The red line represents the point at which $A_{bound} = 2\ A_{fold}$ or where $Affinity_{self} = 0$ as seen in Eq. 5.

A peptide's self-binding affinity includes both the free energy of the self-bound state and its unfolding free energy (Eq. 5). We first examine the relative contributions of these two and, in turn, the sequence characteristics that strongly manipulate self-binding affinity. Figure 4 shows self-bound and folding free energies of all 2080 $(P/H)_{12}$ sequences. A line given by the relation $A_{\text{bound}} = 2\ A_{\text{fold}}$

demarcates the sequences for which the self-binding affinity is zero; on its left, sequences have good self-binding affinity ($\text{Affinity}_{\text{self}} < 0$) while the opposite is true for sequences to the right ($\text{Affinity}_{\text{self}} \geq 0$). Unsurprisingly, the self-bound energy increases with the number of H residues in the sequence since HH contacts give the only favorable intermolecular interaction, aside from the sequence-independent hydrogen bonding term. The relationship is not strictly proportional, however, since both the antiparallel and parallel poses contribute to the free energy (Eq. 2). That being said, parallel binding is usually more preferable and dominates the bound free energy because each H residue aligns with its partner in the other peptide, achieving the maximum HH contacts. Therefore, the degeneracy in A_{fold} values for a sequence of a particular value of A_{bound} is largely due to sequence patterning for the same hydrophobic composition.

On the other hand, the folding free energy bears only weak proportionality to the number of H residues (N_H). Sequences with the same number of H residues (same sequence composition) can exhibit a fairly large range of folding free energies. It is easy to see that composition alone (i.e., the variable N_H) cannot be fully explanatory by noting that H sites separated by an odd number of residues can never form a contact on the 2D lattice. This is a particular feature of the model that places constraints on residue-residue interactions. If all H residues fall on odd- or even-numbered residues such that no HH contacts are possible in any conformation, the folding free energy is always the maximum value, $A_{\text{fold}} = -T \ln \Omega$ where Ω gives the number of distinct conformations, regardless of the number of H residues. Those sequences cluster in the top-right corner of Fig. 4 with $A_{\text{fold}} = -2.89$. Note that if $N_{\text{H}} = 7$, there must be at least one HH contact in any folded conformation. In contrast, the lowest folding free energy is $A_{\text{fold}} = -7.16$ for $N_H = 12$. Therefore, it is clear that sequences to the left of the red line in Figure 4 – those that are good self-binders – are peptides that minimize the stability of their folded monomeric states for a given number of H residues (represented roughly by the x-axis). In other words, good self-binders tend to have less favorable folding free energy, and hence pay a lower cost of unfolding upon dimerization. Chiti and Dobson described a similar effect in the aggregation of globular proteins, noting that the unfolding free energy cost is often anti-correlated with aggregation propensity [22].

The folding free energy thus plays a significant role in self-affinity, and we find that it is well-estimated by the number of hydrophobic residue pairs ($N_{\text{HH pairs}}$) in a sequence. A "HH pair" is defined as two H residues in a sequence that can form a HH contact in at least one conformation; in other words, the pair is separated by an even number of intervening residues. The sequences HPPH, HPPPPH, and HPPPPPPH have HH pairs, while HPH, HPPPH, and HPPPPPH do not due to lattice-geometric constraints. One H residue can be a part of more than one HH pair; for example, there are two HH pairs in a sequence HPPHPH ($N_{\text{HHpairs}} = 2$), although it is also possible that not all HH contacts may be formed simultaneously in any one conformation. The number of HH pairs is useful because it correlates well with the number of possible HH contacts in folded states, and therefore to the folding free energy of a sequence. Table 1 illustrates the relationship between folding free energies and the number of HH pairs for a family of sequences with the same number of H residues, and Figure 5 illustrates how N_{HHpairs} is calculated. The first sequence, with alternating P and H residues, cannot form any HH contacts due to the 2D lattice and its folding free energy is the highest among all the sequences. On the other hand, the last sequence has a repeat HPPH pattern that gives a total of seven HH pairs and particularly low (stable) folding free energy.

Table 1. Sequences with the same number of H residues (N_{H}) have folding free energies that vary widely but that are well-tracked by the number of HH pairs (N_{HHpairs}).

Sequences	N_{H}	$\text{N}_{\text{HHpairs}}$	A_{fold}	A_{bound}	$\text{Affinity}_{\text{self}}$
PHPHPHPHPHPH	6	0	-2.89	-9.60	-3.83
PHPHPHPPPHHH	6	3	-3.35	-9.60	-2.89
PPPPPPHHHHHH	6	4	-4.16	-9.60	-1.28
HPPHHPPHHPPH	6	7	-5.40	-9.60	1.20

Figure 5. Illustration of the number of HH pairs. Each line shows a HH contact possible in at least one configuration on the 2D lattice; these are pairs of H residues separated by an even number of amino acids.

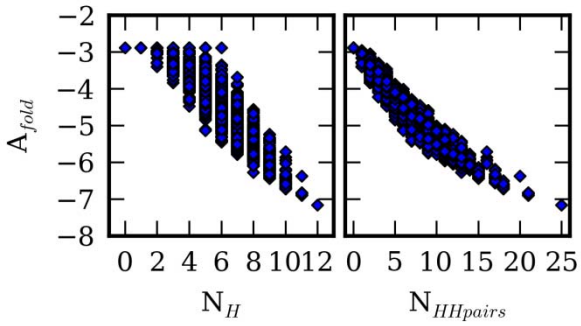

Figure 6. Folding free energies compared to the number of hydrophobic residues and the number of HH pairs, for all 2080 $(P/H)_{12}$ sequences.

Fig. 6 shows that in general the folding free energy is more strongly related to the number of HH pairs in a sequence than the raw number of H residues. At N_H=6, for example, peptides with the same

sequence composition can have folding free energies that vary as large as $8.5k_BT$. On the other hand, the maximum variation in the folding free energy for peptides with identical numbers of HH pairs is only $3.9k_BT$. Therefore, it is not simply the composition, in terms of hydrophobic residues, that determines self-binding propensity, but also the sequence pattern manifested in the distribution of H residues that gives rise to HH pairs. This behavior originates in the unfolding free energy and in particular means that the positions of P residues are important even though they do not interact.

Although this "HH pair" approach cannot be used to quantitatively estimate folding free energies in real peptides or proteins, the physical implication is that the locations of the hydrophilic residues in a sequence can be a significant factor that contributes to unfolding free energy and hence self-binding affinity. In other words, a slight mutational change that introduces, deletes, or substitutes a hydrophilic residue may lead to large changes in the folding free energy because this can modulate the possible interactions between hydrophobic amino acid pairs. This sensitivity is pronounced in peptides because, unlike globular proteins, they often have a diverse conformational ensemble. Here, of course, the 2D lattice model has a unique odd-even effect whereby residues separated by an odd number of amino acids can never make a contact. While this particular behavior is obviously distinct from real proteins, it does emphasize the role that backbone and geometric constraints can have in determining possible intramolecular hydrophobic contacts, which in turn significantly impact unfolding free energies.

B. The impact of unfolding free energies on binding landscapes

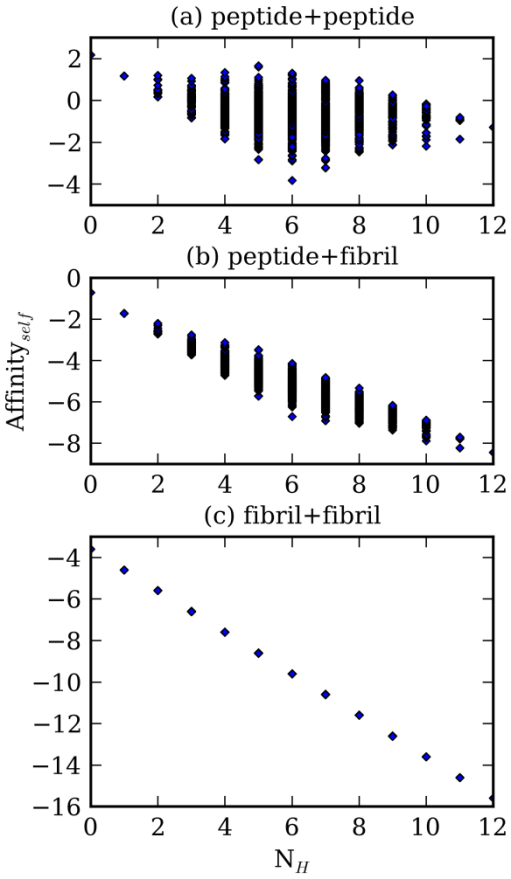

Figure 7. The relationship of the self-binding affinity to the number of H residues (N_H), shown for three scenarios: (a) peptide-peptide binding, (b) peptide addition to a fibril, and (c) association of two fibril ends (see Fig. 1). Points represent every $(P/H)_{12}$ sequence. Panel (a) illustrates that binding between two monomeric peptides is strongly influenced by factors other than sequence composition (i.e., N_H), in contrast to binding events involving fibrils.

Purely energetic pictures of peptide binding emphasize the direct interactions between the peptide and its partner, rather than a full thermodynamic treatment that includes entropies and unbound-state free energies. When is that picture a reasonable representation of the true binding affinities? More specifically, when is the self-binding affinity well-described by the number of H residues of a sequence? Figure 7 examines the correlation of self-binding affinities with N_H for the three distinct scenarios illustrated in Figure 1: (a) dimerization of two peptide monomers, which requires both of their unfolding; binding of a

monomer to a linear peptide already in a fibril, involving unfolding of only one peptide; and binding of two linear peptides each in fibrils, with no unfolding events. As Fig. 7(a) shows, the self-binding affinities of peptide-peptide binding are strongly influenced by factors other than N_H, an effect that originates in the unfolding free energies. On the other hand, the self-binding affinities for both peptide-fibril (Fig. 7 (b)) and fibril-fibril binding (Fig. 7 (c)) show a strong proportionality to sequence composition. Thus, it seems reasonable to approximate the binding affinity using sequence composition and direct interaction energetics when at least one partner undergoes minimal conformational change upon binding.

Despite the simplicity of the lattice model, these results highlight features of peptide-peptide binding that distinguish them from recognition processes involving structured partners; they also emphasize the additional complexity of amyloid assembly over structured protein assembly. Namely, the short length, backbone flexibility, and hence diverse conformational ensemble of peptides in their monomeric state give rise to significant conformational changes upon association that have a profound effect on the sequence-dependence of the binding affinity.

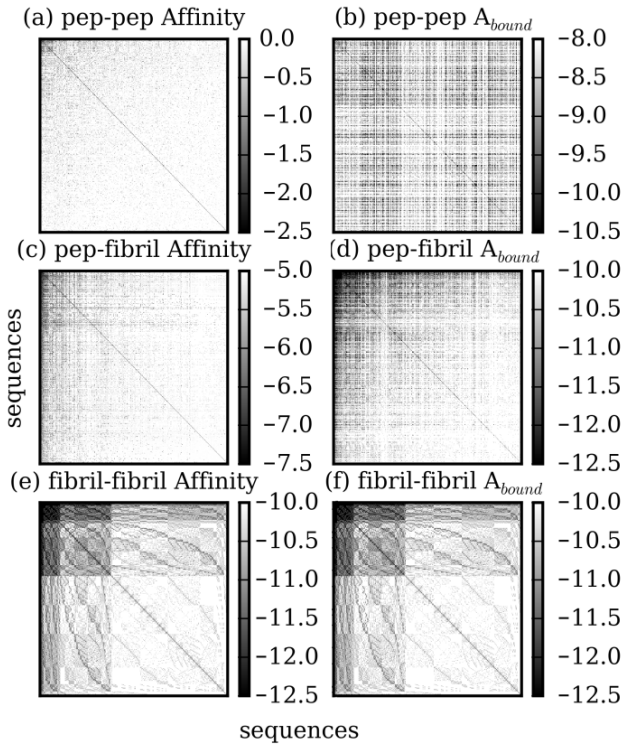

Figure 8. Binding landscapes for the $(H/P)_{12}$ model for the three binding scenarios of Fig. 7. Binding affinities are shown in the left column and bound-state free energies in the right; the former accounts for the unfolding free energies of the peptides, whereas the latter includes only the bound-state partition sum in Eq. 2. Each row/column represents a given sequence, and colors then indicate strength of interaction (lower and more favorable free energies in darker gray). Sequences are sorted by their self-binding affinities (the diagonal lines), so that the most favorable interactions occur in the top left corners and the diagonal in the left panels lightens in color from top-left to bottom-right. The sequence orders are identical among adjacent plots in each row: (a) and (b), (c) and (d), and (e) and (f). After sorting, only the first 400 sequences are shown.

The unfolding free energy for peptide-peptide binding dramatically affects not only self-binding affinity, but the entire binding landscape. Figure 8 illustrates the landscape in a "contact-map" format, where each column/row represents a unique sequence and pixel colors give free energies for the corresponding column-row sequence pair. Fig. 8(a) shows the landscape for the net affinities in peptide-

peptide binding, including the unfolding free energies, while 8(b) shows only the contribution from the bound state and direct interactions. The two landscapes show marked differences, reflecting the critical importance of the unfolding term. On the other hand, the same landscapes for peptide-fibril binding are rather similar, as shown in Figs. 8(c) and 8(d), and are identical by definition in the fibril-fibril binding case, as in Figs. 8(e) and 8(f). To summarize, Fig. 8 shows that sequence effects in peptide-peptide dimerization in this simple model are governed by qualitatively distinct thermodynamic driving forces than oligomer and fibril growth processes. Interestingly, Figs. 8(e) and 8(f) reveal broad areas in the landscape that manifest both homo- and heterophilic binding, particularly as is evident in the dark areas in the top left corners. This loss of specificity and increased binding affinity is suggestive of the finding that amyloid-like fibrils can sequester numerous metastable proteins in cells [23].

C. Unfolding free energies generate binding specificity

To examine truly homophilic binding sequences, it is essential to examine binding specificities in addition to affinities. If a sequence's self-binding affinity is negative and specificity is positive (good affinity and good specificity), the former must be the lowest, most favorable affinity among all possible binding partners; we consider such sequences to be intrinsic homophilic binders. Figure 9 gives a graphical representation of sequence clustering in the space of affinity and specificity, showing that there is a wide range of behavior. A few representative sequences are highlighted with letter codes to show how their behavior varies, over the three binding scenarios; their thermodynamic properties are detailed in Table 2.

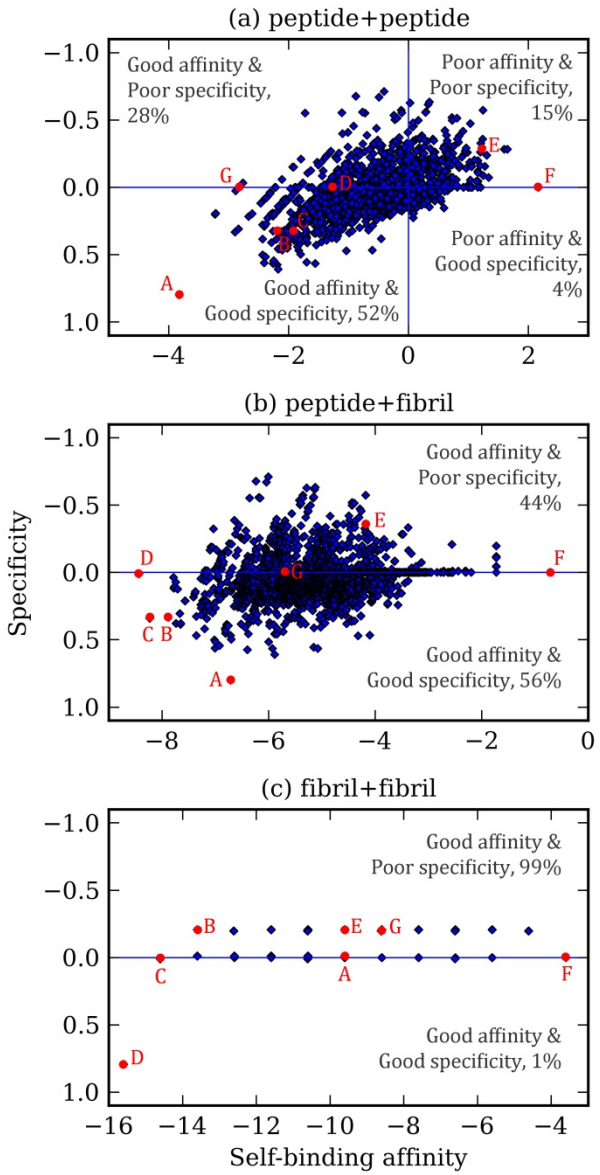

Figure 9. The relationship between self-binding specificities and self-binding affinities of all $(P/H)_{12}$ sequences, for the three binding scenarios of Fig. 1. The y-axis is inverted to present sequences in (a) with both good affinity and good specificity in the third quadrant. Percentages indicate the fraction of sequences in each quadrant.

Figure 9(a) shows for peptide-peptide binding the emergence of families of sequences that appear to fall along common lines in the affinity-specificity space with a slope of 0.5. These lines cluster related

sequences with a similar bound-state free energy and a common partner sequence S' that defines the specificity value (Eq. 7). The variation along these lines is governed primarily by differences in folding free energies within the family, which doubly impact the self-binding affinities and singly the specificity through Eq. 6 – explaining the observed slope and overall positive correlation in Fig. 9(a). On the other hand, self-affinities for peptide-fibril binding are less influenced by folding free energies and the unfolding penalty for binding becomes weaker such that affinities are lower all around, reflected by the global leftward shift in Fig. 9(b). In this case, the free energies of the bound state dominate and most sequences thus have negative self-binding affinities, but interestingly the fraction of sequences with positive specificities remains unchanged. Finally, the case of fibril-fibril binding in Fig. 9(c) shows notably less diverse behavior because folding free energies do not contribute and there are only a few bound-state energy levels (Fig. 3). In this case, the only sequence that has a large positive specificity (point D) is trivially the poly-H sequence, which has the uniquely lowest bound free energy (largest number of H residues) among all sequences.

Table 2. Properties of selected sequences shown in Fig. 9.

	Sequence	A_{bound}	A_{fold}	(a) peptide-peptide		(b) peptide-fibril		(c) fibril-fibril	
				$Affinity_{self}$	Specificity	$Affinity_{self}$	Specificity	$Affinity_{self}$	Specificity
A	PHPHPHPHPHPH	-9.60	-2.89	**-3.83**	**0.80**	-6.71	**0.80**	-9.60	-0.01
B	PHHHHHHHHHHP	-13.60	-5.71	-2.18	0.34	-7.89	0.34	-13.6	-0.21
C	PHHHHHHHHHHH	-14.61	-6.38	-1.85	0.34	-8.23	0.34	-14.6	0.01
D	HHHHHHHHHHHH	-15.60	-7.16	-1.28	0.01	**-8.44**	0.01	**-15.6**	**0.79**
E	HPPHHPPHHPPH	-9.60	-5.40	1.21	-0.33	-4.20	-0.33	-9.60	-0.21
F	PPPPPPPPPPPP	-3.60	-2.89	2.17	0.00	-0.71	0.00	-3.60	0.00
G	PPPHPHPHPHPH	-8.60	-2.89	-2.83	0.00	-5.71	0.00	-8.60	-0.21

Table 2 illustrates the thermodynamic properties that govern the representative sequences shown in Fig. 9. The alternating P/H sequence has the lowest self-binding affinity and is unique because the bound-state free energy is moderate with six H sites, but the unfolding free energy cost is very small since it can never form HH contacts – a consequence of the 2D lattice constraints. (Fig. 6 underscores that sequences with up to six hydrophobic residues can still achieve small unfolding free energies.) By comparison, alternating polar and hydrophobic sequence patterns in real peptides are thought to emerge instead due to the orientation and co-location of hydrophobic side chains on one side of the beta sheet geometry, an effect that is absent in the lattice model. Thus the special behavior of the alternating P/H sequence in the model is influenced quantitatively by the 2D lattice. Nonetheless, the model still emphasizes the role of the monomeric state in achieving affinity, and the potential for backbone configurational constraints to influence this state's free-energetic stability. Real peptides experience a distinct set of backbone geometric constraints, but these will likely still influence the kinds of intramolecular contacts and hence sequence patterns that will lower the unfolding free energy cost.

We find other patterns of hydrophobic amino acids that give rise to unique self-binding properties. Sequence B has both good self-binding affinity and specificity, and involves a hydrophobic core flanked by hydrophilic ends, a pattern observed in amyloidogenic sequences [24,25]. In contrast, the all-H sequence (D) naturally has good self-binding affinity, but almost no specificity because many other sequences bind with similar affinities. Its binding affinities also suffer from a particularly large unfolding free energy cost. A sequence that can form alpha-helical conformation on a 2D lattice (E) also has a high unfolding free energy and thus has a poor self-binding affinity and a poor specificity. The last sequence highlighted (G) is distinct from the alternating H/P sequence (A) through a single H→P mutation; surprisingly, the addition of a single hydrophilic residue completely eliminates the specificity of this sequence because it now binds to A with equivalent affinity as to itself.

In the case of peptide-fibril binding (Fig. 9 (b)), the bound-state free energy becomes the important contributor to self-binding affinity; sequences with large numbers of hydrophobic residues shift

to the left side of the plot relative to their positions in Fig. 9 (a). Here, the sequence with all hydrophobic residues (D) has the lowest self-binding affinity, although its specificity remains as bad as in Fig 9 (a). In the case of fibril-fibril binding (Fig 9 (c)), the sequence with all hydrophobic residues (D) still has the lowest self-binding affinity and this sequence is the only one with positive specificity due to its exceptionally low self-binding free energy.

D. Homophilic sequences are enriched in clustered hydrophobic residues

To understand the representative sequence characteristics of each quadrant of Fig 9 (a), we use a string parsing technique that computes the frequency of words (strings of characters of a given length) within a sequence [26,27]. For example, the sequence HPPPH has three two-words, HP (25% frequency), PP (50%), and PH (25%). We find that the frequency of HHH words, f(HHH), shows clear differences between each quadrant, as shown in Table 3. Namely, the HHH pattern is enriched by a factor of 50% for sequences with both good self-binding affinities and specificities, it reduced for the other three cases. This is related to the fact that there are slightly more H residues for the former sequences, as shown in Table 4.

Table 3. Frequencies of three letter words of sequences in each quadrant of Fig 9 (a)

self affinity	self specificity	f(PPP)	f(PPH)	f(PHP)	f(PHH)	f(HPP)	f(HPH)	f(HHP)	**f(HHH)**
good	good	0.12	0.13	0.08	0.17	0.09	0.10	0.12	**0.18**
good	poor	0.12	0.14	0.17	0.12	0.11	0.15	0.09	**0.09**
poor	good	0.08	0.17	0.15	0.15	0.14	0.15	0.11	**0.05**
poor	poor	0.16	0.18	0.17	0.10	0.14	0.13	0.08	**0.04**

Table 4. Average properties of sequences in each quadrant from Fig.9 (a)

self affinity	self specificity	N_P	N_H	A_{bind}	A_{fold}	$Affinity_{self}$	Specificity	$N_{HHpairs}$
good	good	5.61	6.39	-6.39	-4.49	-1.02	0.15	6.67
good	poor	6.25	5.75	-5.75	-4.31	-0.73	-0.07	5.64
poor	good	6.08	5.92	-5.92	-4.88	0.24	0.07	7.13
poor	poor	6.82	5.18	-5.18	-4.61	0.43	-0.12	5.87

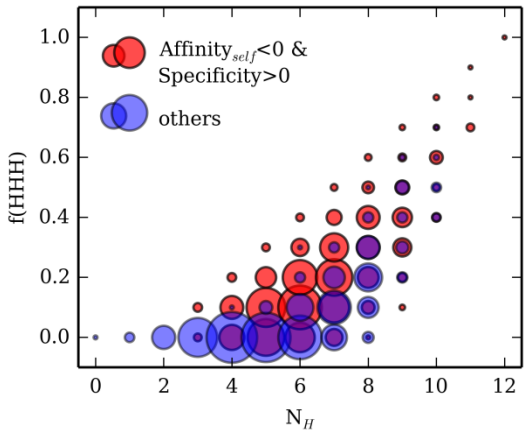

Figure 10. Frequency of triple consecutive hydrophobic residues in a sequence, f(HHH), with respect to the number of hydrophobic residues, N_H. Sequences with both good affinity and specificity for the peptide-peptide scenario (as defined in Fig 9a) are shown as red circles, while all other sequences are shown as blue circles. The area of each circle is proportional to the number of sequences at each point.

Figure 10 shows f(HHH) as a function of N_H for sequences with good peptide-peptide self-binding affinities and good specificities (red), compared to that of all other sequences (blue). Both groups span various N_H and f(HHH) values, and show that sequences with larger number of H residues have a higher frequency of HHH words. However, it is clear that sequences with both good affinity and

specificity maximize the frequency of HHH words for a given sequence composition. In other words, clustering of H residues into hydrophobic patches (like sequences B and C in Fig. 9) minimizes the number of HH pairs and hence the unfolding free energy penalty; this improves binding affinity all around but the effect is doubly pronounced for the *self*-binding affinity and hence it raises specificity is high as well. This observation is consistent with a 'hydrophobic patch' or 'hydrophobic hot spot' model of aggregation in amyloidogenic peptides; indeed, frequencies of consecutive hydrophobic residues are lower in soluble proteins [28]. Here, our results suggest a slightly distinct interpretation in terms of the role that the unbound state and hence unfolding free energy plays.

E. Homophilic binding of peptides with charged ends

A natural question is the role that charged amino acids play in the self-binding picture, since it is thought that such moieties contribute strongly to fibril formation in many amyloidogenic peptides [29,30], and several recent theoretical efforts have sought to understand the role of sequence charge patterning on the properties of intrinsically disordered peptides [31,32]. Here, we examine sequences with charged amino acids at the termini, rather than at any arbitrary location, in order to mitigate the significantly increased sequence space. Naturally, uncapped peptides possess oppositely-charged termini at neutral pH, and the modulation of their charged states is known to significantly affect their fibrillization even though the change in their net charge is small [24,33].

In addition to H and P monomer types, we introduce C (cationic, positively charged) and A (anionic, negatively charged) amino acids. Table 5 gives the interaction matrix between all type pairs; we take the simplest approach and introduce interactions only for AA, CC, and AC contacts, which have the same magnitude as HH energetics. To compare with the results above using N=12, we compute self-binding affinities and specificities of sequences with 10 central P/H residues and P/H/C/A residues at the ends, for peptide-peptide binding only.

Table 5. Contact energy matrix between P (hydrophilic), H (hydrophobic), C (cationic), and A (anionic) residues. All energies are dimensionless.

	P	H	C	A
P	0	0	0	0
H	0	-1	0	0
C	0	0	+1	-1
A	0	0	-1	+1

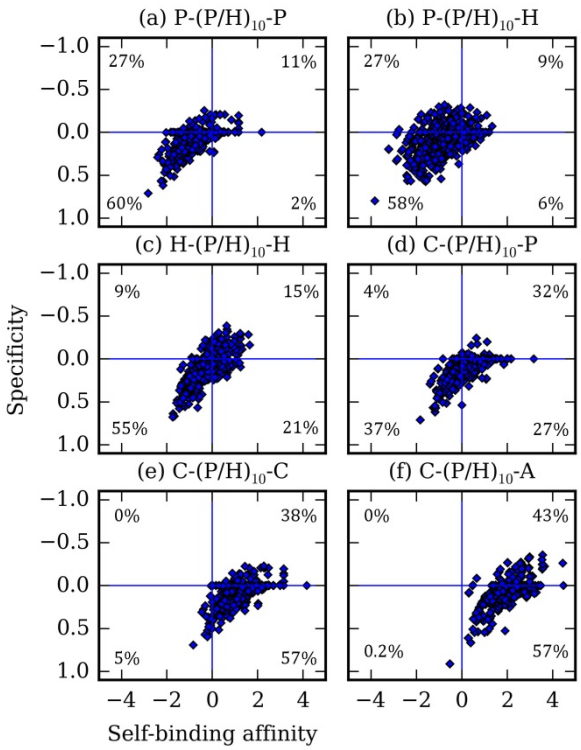

Figure 11. Self-binding affinity and specificity in peptide-peptide binding for the sequence space (A/C/H/P)-$(H/P)_{10}$-(A/C/P/H), that is, for a modified HP model in which there are four different types of end residues. Percentages indicate the fraction of the sequence space falling within each quadrant.

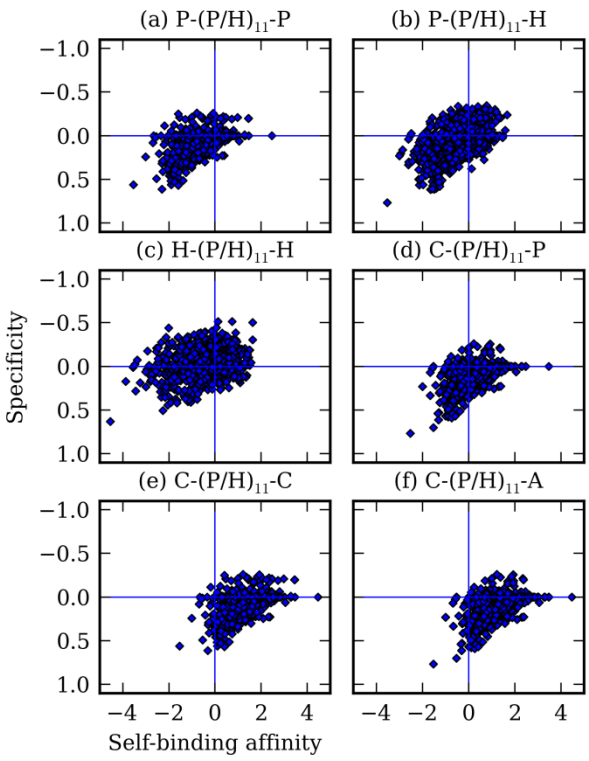

Figure 12. Self-binding affinity and specificity plots for sequences with distinct termini chemistry and an odd length (N=13).

Figure 11 illustrates the behavior of the affinity-specificity landscape for all (A/C/H/P)-$(H/P)_{10}$-(A/C/P/H) peptides in response to the chemistry at the ends of the chains. Notably, the ranges of affinities in panels (a)-(f) show variations. When both ends are hydrophobic (c), the self-binding affinity shifts slightly to right (weaker affinity) compared to the case in which the termini are completely polar (a) or involve both H and P (b). This occurs because the two terminal hydrophobic residues in a sequence where N is even can always form an HH contact (for example, in a hairpin-like configuration) and thus such sequences are likely biased towards a higher unfolding cost. In case of odd sequence length (N=13), contact between the termini is impossible and the shift becomes negligible, as shown in Figure 12.

The addition of terminal charge generally weakens self-binding affinity. Both cases involving a single charged residue at one end (panel d) and two like-charged residues at the termini (e) weaken binding affinity due to charge-charge repulsions that occur in at least one of the orientations (parallel or anti-parallel) of the binding pose. Interestingly, the presence of oppositely charged residues at the termini reduces the self-binding affinity the most (f), even with the possibility to form stabilizing salt-bridge-type interactions in the anti-parallel binding pose. The reason is that the same interactions significantly destabilize the parallel pose, which is always the more favorable alignment because the sequences align so as to maximize the number of interpeptide HH contacts.

In contrast, the specificities are less affected by the chemistry at the termini. In all cases (panels a-f), the alternating P/H sequence still has amongst the highest hemophilic specificity. These conclusions persist for peptides with both odd and even numbers of residues.

IV. DISCUSSIONS AND CONCLUSIONS

This work addresses the question: what do the binding landscapes of short peptides look like, across all of sequence space for a given length, in a simple lattice model that allows exhaustive exploration of it? The binding landscape compares the intrinsic binding affinities of all sequences with all other sequences, and in turn, includes information that describes the intrinsic homophilic specificities of these interactions, i.e., the extent to which a peptide will preferentially bind to itself in the face of many other potential partners differing in sequence. We find several emergent features in the model binding landscapes that, while not quantitatively representative of real peptides, nevertheless suggest ways in by which complexity can arise due to quite general features of peptide molecular physics.

For example, we find that the sequence with alternating hydrophobic and hydrophilic residues has the best self-binding affinity and specificity in this model because it limits potential hydrophobic-hydrophobic contacts in the monomeric state, which create a high unfolding free energy cost upon

binding. This particular motif is known to favor the formation of beta-sheet secondary structures in real proteins for distinct reasons [15,34], but interestingly, patterns of alternating hydrophobic and hydrophilic residues occur less frequently in natural proteins than would be expected based on typical amino acid compositions [35]. The inability to form stabilizing contacts of hydrophobic residue pairs that are close to each other in sequence is key in the model. The 2D lattice constraints magnify these effects, and indeed residues separated by odd numbers of amino acids can never form a contact, no matter how distant in sequence. However, real proteins still see constraints on interactions of sequence-close animo acids, due to the geometry of permissible backbone arrangements. Thus one still expects limitations on favorable local amino acid interactions, which may play out in qualitatively similar ways.

We also find that sequences with clustered hydrophobic residues have both good binding affinities and specificities, using a word frequency analysis on the subset of sequences with such properties. Statistical studies of natural protein sequences by King and coworkers [28,36] have also revealed that groups of three or more hydrophobic residues occur less frequently than would be expected assuming neutral selection, suggesting that clusters of hydrophobic residues have been selected against during protein evolution.

Lastly, we find that a dimerization process involving binding between two free peptides manifests significant differences from the oligomer/fibril growth processes or protein-protein binding processes. The cost, in terms of free energy, for "unfolding" the peptides from an intrinsic, potentially diverse conformational ensemble in the monomeric state to a structured binding pose has profound effects on the entire sequence binding landscape for peptide dimerization. Indeed, while a sequence's hydrophobicity dictates the direct binding interactions, it has a less predictable effect on the thermodynamics of the unbound conformational ensemble, such that it correlates poorly with the overall free energy of the binding process. The importance of peptide unfolding in fibrillation was also highlighted by Uversky and Fink [1], who noted that most amyloidogenic proteins need to overcome considerable conformational rearrangement for fibrillization occur. They suggested that amyloidogenic intermediates have mostly pre-

molten globule conformations, which reduces entropic costs upon assembly. Furthermore, Rauscher et al. [37] found that amyloidogenic sequences generally lack glycine residues, which are conformationally plastic such that they increase unfolding costs.

In this study, we examined all sequences irrespective of concentration because sequence motifs with high homophilicity are likely to be intrinsically down-regulated in the cellular environment due to high aggregation propensity, or otherwise evolutionarily disfavored. Indeed, Tarataglia et al. detected a strong anti-correlation between human protein expression levels and measured aggregation rates [38]. However, the model presented here could be readily extended to address concentration effects. A simple binding between two sequence S and S' follows,

$$S + S' \rightarrow SS'. \tag{14}$$

The binding affinity gives the standard free energy of reaction and association constant:

$$K = \exp[-\Delta G^{\circ}] = \frac{[SS']}{[S][S']} \tag{15}$$

where the brackets as usual indicate species concentrations. Using the notation introduced earlier, we have

$$\text{Affinity}\,(S, S') = \Delta G^{\circ} = -T \ln\,[SS'] + T\,\ln[S] + T\,\ln[S']. \tag{16}$$

Thus we can define an *effective* binding affinity that incorporates concentration as

$$\text{Affinity}_{\text{eff}}(S, S') = -T \ln\,[SS'] = \text{Affinity}(S, S') - T\,\ln[S] - T\,\ln[S']. \tag{17}$$

Using the definition introduced in Eq. 6, the specificity becomes proportional to the log of the S to S' concentration ratio:

$$\text{Specificity}_{\text{eff}}(S) = \text{Affinity}(S, S') - \text{Affinity}_{\text{self}}(S) + T\,\ln\frac{[S]}{[S']}. \tag{18}$$

In particular, if the concentration of S is greater than any other species, then its effective homophilicity is enhanced. Unsurprisingly, then, a prominent population of one sequence has the potential to induce homophilic binding. A more realistic analysis would seek to estimate in vivo concentrations on the basis of sequence, and thus in turn, the effective binding landscape. For example, one might estimate protein hydrophobicity from sequence using the Eisenberg hydrophobicity consensus scale [39], and correlate this with observed sequence expression rates

Furthermore, here we studied sequences made of predominately two types of residues (hydrophobic and polar) and considered mainly attractive inter-reside interactions. However, if there are more diverse interactions, as in real proteins and peptides with 20 amino acids, it may be easier to generate specificity due to the increased means for tuning the binding energy with respect to the unfolding free energy. For example in *Drosophila melanogaster,* the down syndrome cell-adhesion molecule (Dscam) proteins, which function as molecular tags to regulate neuronal connectivity in the fly brain, have 38016 isoforms with 19008 extracellular domains that act as surface receptors [40,41]. Each isoform shows highly exclusive isoform-specific homophilic binding; 95% of all isoforms exclusively bind to another copy of the identical isoform. The molecular mechanism that underlies the homophilic-binding specificity is still unknown. A consideration of the binding landscapes in highly sequence diverse peptides, using simple models like the one investigated here, may provide an approach towards understanding homophilic binding in this and related problems.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the Alfred P. Sloan Foundation and the National Science Foundation (Project DMR-1312548).

REFERENCES

[1] V. N. Uversky and A. L. Fink, Biochim. Biophys. Acta BBA - Proteins Proteomics **1698**, 131 (2004).

[2] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **75**, 333 (2006).

[3] U. Moran, R. Phillips, and R. Milo, Cell **141**, 1262 (2010).

[4] M. R. H. Krebs, L. A. Morozova-Roche, K. Daniel, C. V. Robinson, and C. M. Dobson, Protein Sci. **13**, 1933 (2004).

[5] A. R. A. Ladiwala, M. Bhattacharya, J. M. Perchiacca, P. Cao, D. P. Raleigh, A. Abedini, A. M. Schmidt, J. Varkey, R. Langen, and P. M. Tessier, Proc. Natl. Acad. Sci. **109**, 19965 (2012).

[6] J. M. Perchiacca, A. R. A. Ladiwala, M. Bhattacharya, and P. M. Tessier, Proc. Natl. Acad. Sci. **109**, 84 (2012).

[7] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, and C. M. Dobson, Nature **424**, 805 (2003).

[8] S. Maurer-Stroh, M. Debulpaep, N. Kuemmerer, M. L. de la Paz, I. C. Martins, J. Reumers, K. L. Morris, A. Copland, L. Serpell, L. Serrano, J. W. H. Schymkowitz, and F. Rousseau, Nat. Methods **7**, 237 (2010).

[9] M. Belli, M. Ramazzotti, and F. Chiti, EMBO Rep. **12**, 657 (2011).

[10] M. H. Hecht, A. Das, A. Go, L. H. Bradley, and Y. Wei, Protein Sci. **13**, 1711 (2004).

[11] A. Trovato, F. Chiti, A. Maritan, and F. Seno, PLOS Comput Biol **2**, e170 (2006).

[12] K. A. Dill, Biochemistry (Mosc.) **29**, 7133 (1990).

[13] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[14] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).

[15] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht, Science **262**, 1680 (1993).

[16] M. W. West, W. Wang, J. Patterson, J. D. Mancias, J. R. Beasley, and M. H. Hecht, Proc. Natl. Acad. Sci. **96**, 11211 (1999).

[17] W. Kim and M. H. Hecht, Proc. Natl. Acad. Sci. **103**, 15824 (2006).

[18] A. W. Fitzpatrick, T. P. J. Knowles, C. A. Waudby, M. Vendruscolo, and C. M. Dobson, PLoS Comput Biol **7**, e1002169 (2011).

[19] N. London, D. Movshovitz-Attias, and O. Schueler-Furman, Structure **18**, 188 (2010).

[20] S. J. Fleishman and D. Baker, Cell **149**, 262 (2012).

[21] G. Grigoryan, A. W. Reinke, and A. E. Keating, Nature **458**, 859 (2009).

[22] F. Chiti and C. M. Dobson, Nat. Chem. Biol. **5**, 15 (2009).

[23] H. Olzscha, S. M. Schermann, A. C. Woerner, S. Pinkert, M. H. Hecht, G. G. Tartaglia, M. Vendruscolo, M. Hayer-Hartl, F. U. Hartl, and R. M. Vabulas, Cell **144**, 67 (2011).

[24] M. López de la Paz, K. Goldie, J. Zurdo, E. Lacroix, C. M. Dobson, A. Hoenger, and L. Serrano, Proc. Natl. Acad. Sci. U. S. A. **99**, 16052 (2002).

[25] M. L. de la Paz and L. Serrano, Proc. Natl. Acad. Sci. **101**, 87 (2004).

[26] O. Bonham-Carter, J. Steele, and D. Bastola, Brief. Bioinform. bbt052 (2013).

[27] S. Vinga and J. Almeida, Bioinformatics **19**, 513 (2003).

[28] R. Schwartz, S. Istrail, and J. King, Protein Sci. **10**, 1023 (2001).

[29] S. Zhang and A. Rich, Proc. Natl. Acad. Sci. **94**, 23 (1997).

[30] J. Zurdo, J. I. Guijarro, J. L. Jiménez, H. R. Saibil, and C. M. Dobson, J. Mol. Biol. **311**, 325 (2001).

[31] R. K. Das and R. V. Pappu, Proc. Natl. Acad. Sci. **110**, 13392 (2013).

[32] L. Sawle and K. Ghosh, J. Chem. Phys. **143**, 085101 (2015).

[33] J. Jeon and M. S. Shell, Biophys. J. **102**, 1952 (2012).

[34] M. W. West and M. H. Hecht, Protein Sci. **4**, 2032 (1995).

[35] B. M. Broome and M. H. Hecht, J. Mol. Biol. **296**, 961 (2000).

[36] R. Schwartz and J. King, Protein Sci. **15**, 102 (2006).

[37] S. Rauscher, S. Baud, M. Miao, F. W. Keeley, and R. Pomès, Structure **14**, 1667 (2006).

[38] G. G. Tartaglia, S. Pechmann, C. M. Dobson, and M. Vendruscolo, Trends Biochem. Sci. **32**, 204 (2007).

[39] D. Eisenberg, R. M. Weiss, T. C. Terwilliger, and W. Wilcox, Faraday Symp. Chem. Soc. **17**, 109 (1982).

[40] W. M. Wojtowicz, J. J. Flanagan, S. S. Millard, S. L. Zipursky, and J. C. Clemens, Cell **118**, 619 (2004).

[41] W. M. Wojtowicz, W. Wu, I. Andre, B. Qian, D. Baker, and S. L. Zipursky, Cell **130**, 1134 (2007).