PROTEINS: Structure, Function, and Genetics 50:552–562 (2003)

Helix Propensities of Short Peptides: Molecular Dynamics Versus Bioinformatics

Christopher Bystroff[1]* and Shekhar Garde[2]
[1]*Department of Biology, Rensselaer Polytechnic Institute, Troy, New York*
[2]*Department of Chemical Engineering, Rensselaer Polytechnic Institute, Troy, New York*

ABSTRACT **Knowledge-based potential functions for protein structure prediction assume that the frequency of occurrence of a given structure or a contact in the protein database is a measure of its free energy. Here, we put this assumption to test by comparing the results obtained from sequence-structure cluster analysis with those obtained from long all-atom molecular dynamics simulations. Sixty-four eight-residue peptide sequences with varying degrees of similarity to the canonical sequence pattern for amphipathic helix were drawn from known protein structures, regardless of whether they were helical in the protein. Each was simulated using AMBER6.0 for at least 10 ns using explicit waters. The total simulation time was 1176 ns. The resulting trajectories were tested for reproducibility, and the helical content was measured. Natural peptides whose sequences matched the amphipathic helix motif with greater than 50% confidence were significantly more likely to form helix during the course of the simulation than peptides with lower confidence scores. The sequence pattern derived from the simulation data closely resembles the motif pattern derived from the database cluster analysis. The difficulties encountered in sampling conformational space and sequence space simultaneously are discussed. Proteins 2003;50:552–562.** © **2003 Wiley-Liss, Inc. Key words:**

Key words: protein folding; I-sites library; motifs; secondary structure propensity; structure prediction; knowledge-based potentials; template effect

INTRODUCTION

How is a database like a force field? It sounds like a riddle, but knowledge-based empirical energy functions make an implicit assumption that the database is, in some sense, an equilibrium sample of states. We assume this when we derive "energies" from database counting statistics. By directly comparing a physics-based force field to one derived from database statistics, we can test this assumption.

The helix propensity of a short sequence may be derived from the statistics of sequence-structure correlations, either using neural nets, profiles, or hidden Markov models.[1–6] These methods have been shown to predict helices accurately in short windows of protein sequences. This suggests that the energetic stability of the helix is intrinsic to the helix sequence, and does not come from tertiary interactions.

However, helices in proteins are mostly amphipathic, with a characteristic "heptad repeat" sequence pattern.[7] One side of the helix has hydrophobic side-chains that pack against another hydrophobic surface on the protein, such as another helix or a sheet. The tertiary hydrophobic interactions would stabilize the helical structure even if it were not stable on its own. We may call this the "template effect," similar to the concept of "induced fit" in substrate binding. Either the template effect or intrinsic helicity would equally well explain the success of knowledge-based helix prediction methods, because the proteins used to validate the methods generally also have the hydrophobic template. Figure 1 illustrates the template concept by showing the location of the surrounding protein relative to 30 superimposed amphipathic helices, chosen at random.

Can a knowledge-based method predict helix propensity in short peptides, where there is no template effect? Structures of short peptides, and especially peptides that do not form a stable structure, are not readily available, despite considerable effort using nuclear magnetic resonance (NMR[8–12]). Therefore, to answer this question, we compare the knowledge-based helix predictions of the I-sites library[3,13] with helix content estimates derived from explicit molecular dynamics (MD) simulations, performed using AMBER.[14,15]

AMBER uses pairwise atom–atom distance-based energy functions, including harmonic restraints on covalent geometry, torsion angle potentials, van der Waals interactions, and Coulomb's law to simulate the trajectory of a peptide in explicit aqueous solvent. We may estimate helix propensity by counting the occurrence of helix in a long simulation of a fully solvated peptide. As the length of the simulation increases, so does the reproducibility of the propensity measure. Several recent articles have reported simulation studies of short peptides over timescales relevant to the peptide folding process that are also beginning to overlap with the timescales of fast experimental methods.[16–20]

*Correspondence to: Christopher Bystroff, Department of Biology, Rensselaer Polytechnic Institute, Troy, NY 12180. E-mail: bystrc@rpi.edu

Received 12 October 2001; Accepted 24 July 2002

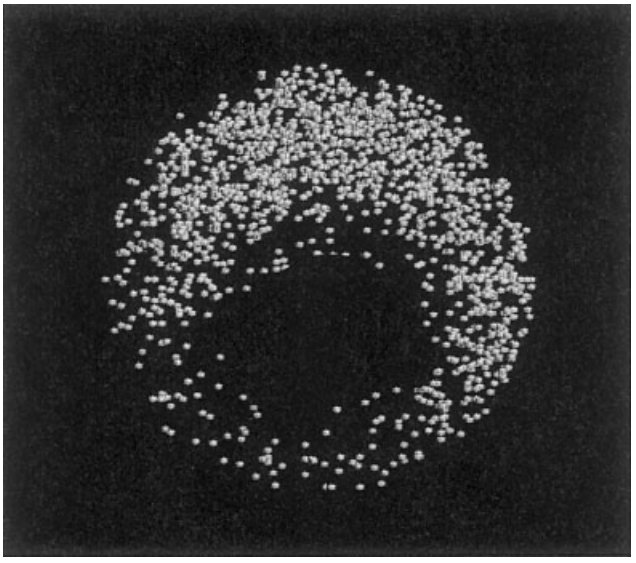

Fig. 1. The database-averaged environment of an amphipathic helix. Dots represent the positions of non-local alpha carbons relative to an amphipathic helix. One hundred examples of amphipathic helix were superimposed with the nonpolar side up. This illustrates the conserved, nonspecific interactions that may have a role in inducing helix formation.

I-sites is a library of sequence motifs, described as 20-by-L log-likelihood matrices, where L is the length of the motif ($3 \leq L \leq 19$). Each motif is derived by summing the frequencies of each of the 20 amino acids at each of the L positions over a large set of protein fragments with the same three-dimensional structures. Using the motif for eight-residue amphipathic helix (I-sites motif 8055), we may estimate helix propensity of an eight-residue peptide sequence by summing the log-likelihood score. I-sites predictions have been validated in "CASP" experiments,[13] and several short peptides with known solution structures fold to the conformation predicted by I-sites.[9,10,21–25] The structure of a seven-residue I-sites "diverging turn" motif has been solved by NMR[12] and was also found to be stable using free-energy calculations.[26]

Knowledge-based and physics-based potentials have been previously compared in their application to the fold recognition problem.[27,28] But in this arena, the physics-based potentials have a distinct disadvantage, because no simulation of the equilibrium distribution of conformational states is possible. Detailed simulations using physics-based potentials such as AMBER or CHARMm[29] are limited to fairly short peptides, the longest attempted being 36 residues.[30]

In this study, we selected 64 eight-residue peptide sequences from proteins of known structure. About half of these sequences match the amphipathic helix motif with a high score. If the intrinsic stability hypothesis holds true, then the high-scoring group should be more helical in the MD simulations than the low-scoring group.

We are aware of many difficulties involved in performing meaningful comparisons of the sort proposed here. Both MD and bioinformatics have their limitations. For example, the validity of classical force fields used in MD simulations for peptide structure predictions as well as the ability of conventional MD simulations to sample conformation space has not been firmly established. Statistical approaches, however, are computationally efficient but do not have access to the underlying physical sources of the statistics. Sparse sampling of sequence space means that only general patterns may be discovered, and we are forced to assume that different positions contribute independently (additively) to the "energy." The currently used knowledge-based approaches are imperfect[31] and continue to be developed.

Despite the difficulties, we find that there is much to be learned from combined MD/bioinformatics approaches. We comment on ways to overcome the aforementioned difficulties and present suggestions for improving the existing I-sites library for better peptide local structure predictions. Connecting the protein database with the fundamental forces of nature may eventually tell us something about how proteins fold.

METHODS

The Sequence and Structure Database

The database for the development of the I-sites library, including the motif studied in this work, consists of 471 protein sequence families from the HSSP database.[32] Each sequence family contains a known structure from the Protein Data Bank (PDB).[33] These 471 families are a subset of the PDBSelect25 list of non-redundant sequence families[32] (October 1997 release), having no more than 25% sequence identity between any two families. Families in the PDBSelect25 list were excluded if the parent structure was not well determined, if the protein was membrane-bound, or if it contained a large number of disulfide bonds. Disordered loops were omitted. Gaps and insertions in aligned sequences were ignored.

The I-sites Library

The I-sites library is a collection of short sequence patterns that correlate with local structure.[3] They are "motifs" in the sense that they map sequences to structures. The motifs were learned from the database using an iterative clustering/reenforcement learning approach. Included in the library are a total of 262 sequence patterns for beta-turns, bulges, half-turns, loops, strands, helix caps, and alpha helix. Sometimes more than one sequence motif maps to the same structure type. In this study, we discuss only the amphipathic helix motif with the I-sites identifier 8055. For more information, see http://isites.bio.rpi.edu.

Peptide Selection

All-eight-residue sequence segments from the non-redundant database of known protein structures were scored versus the motif 8055, as described below. The 64 sequences selected for study were chosen at random from this list, with a bias toward the higher-scoring peptides, such that both extremes of the confidence range (0–1.0) are well represented.

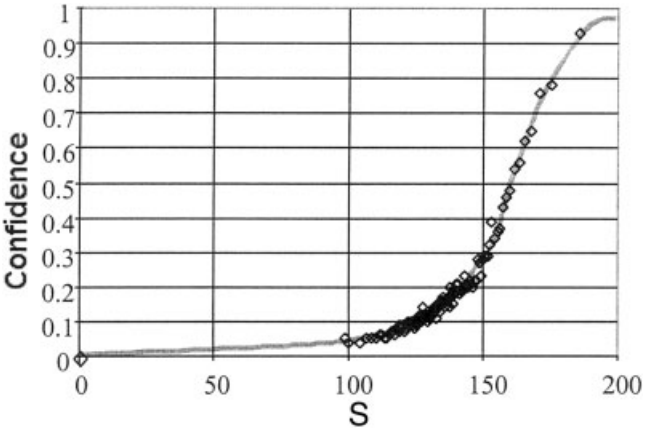

Fig. 2. Confidence curve for I-sites motif 8055. The top-scoring 52000 sequence profile segments of length 8 in the dataset that was used for training the I-sites library were sorted by score, and separated into bins of 500. Each point represents the average score (see Eq. 1) of one bin (S) versus the fraction of those 500 peptides whose backbone angles matched the helix motif (*confidence*).

Estimation of I-Sites Confidence for Helix Formation

The method for obtaining an I-sites score has been described previously,[3] and is repeated here for clarity.

The I-sites confidence score for amphipathic helix is calculated in four steps. First, the sequence of the protein is aligned to its homolog sequences using Psi-Blast.[34] Then, the multiple sequence alignment is condensed to an amino acid probability profile,[35] using sequence weights to correct for evolutionary biases in the homolog set.[36] Third, each eight-residue window of the profile P^T is scored versus the motif profile P^M using the following equation:

$$S(t) = \sum_{i=0,7} \sum_{a=1,20} \log\left(\frac{P^T_{(t+i)a} + \alpha F_a}{(1+\alpha)F_a}\right)\log\left(\frac{P^M_{ia} + \alpha F_a}{(1+\alpha)F_a}\right) \quad (1)$$

where P_{ia} is the probability of amino acid a at position i. Superscript T denotes the target sequence, t is the start of the eight-residue window in T, and superscript M denotes the I-sites scoring matrix for motif 8055. F_a is the overall frequency of amino acid a in the database. α is a "pseudocount" used to buffer the effect of low counts. α = 0.5 was used.

Finally, the score *S(t)* is mapped to a *confidence* using a precalculated histogram of scores for known proteins, plotted against the fraction helical. Figure 2 shows a plot of points representing bins of similar-scoring sequences, their average score versus percent helix. This curve maps each score (S) to a *confidence*.

Note that the information used in this scoring method normally comes from a sequence family, not from a single sequence. Implied in this approach is the idea that all homologs have the same local structure in the aligned positions. This is a common assumption in deriving knowledge-based potentials, and will be discussed later.

Simulation Details

MD simulations were performed at constant temperature (340 K) and pressure (1 atm) for a solution containing one peptide, solvated by approximately 850 TIP3P water molecules and sodium or chloride ions required to neutralize the charge on the peptide. Peptides were N-terminally acetylated and C-terminally methyl-amidated, or uncapped, as indicated under "capped" in Table 1. Simulations were performed using AMBER6.0[14,15] with the "parm94" forcefield.[37] The simulation temperature, 340 K, was found to maximize conformational sampling without unfolding helices. At lower temperatures (280–320 K), we observed high barriers to backbone conformational shifts.

Periodic boundary conditions were applied, and electrostatic interactions were calculated using the particle mesh Ewald (PME) method.[38] Bonds involving hydrogens were constrained using the SHAKE algorithm[39] and Berendsen's coupling algorithms were used to maintain constant temperature and pressure.[40] Either extended or a random initial configuration was equilibrated for at least 1.0 ns. Productions runs of 10 to 30 ns were then performed, and configurations were stored every picosecond for further analysis.

Clustering of Trajectories

Clustering was done to evaluate the sampling of conformational space. Conformations of a given peptide in a molecular dynamics trajectory were clustered using a modified "greedy" approach. Two metrics were used to assess the similarity between any two peptide conformations: first, the maximum difference between backbone phi/psi angles (*mda*) over the central six residues,

$$mda = \max_{i=2,7} (\Delta\phi_i, \Delta\psi_i) \quad (2)$$

and second, the root-mean-square difference between internal Cα–Cα distances, or "distance matrix error" (*dme*):

$$dme = \sqrt{\frac{\sum_{i=4}^{8}\sum_{j=1}^{i-3}(d^{s1}_{ij} - d^{s2}_{ij})^2}{15}} \quad (3)$$

where d^{s1}_{ij}, d^{s2}_{ij} are the distances between alpha-carbons i and j in conformations $s1$ and $s2$, respectively. The *dme* and *mda* were calculated between each pair of configurations and a cutoff was applied. Pairs of configurations were linked by an edge if both metrics fell below their respective threshholds (*dmecut, mdacut*); otherwise they were unlinked. The configuration with the greatest number of edges was selected as the center of the first cluster, and every configuration with an edge to the center was placed into a cluster. The cluster was removed, and the process was repeated until no configurations remained.

The combination of the two metrics (*dme* and *mda*) was found to be the best descriminator for finding correlations between sequence and structure.[3] The cutoff values (*dmecut, mdacut*) were refined by inspecting the clusters. The cutoff values were raised if two clusters seemed to be too similar, or lowered if a single cluster was found to contain two or more dissimilar conformations. The optimal cutoff values, found by trial and error, were *dmecut* = 1.3 Å and *mdacut* = 60°.

TABLE I. Sixty-Four Peptides Studied by MD, Sorted by Confidence[†]

Peptide sequence	I-sites confidence	Helicity	Length of trajectory (ns)	Helix in context?	Capped?
1 YESHVGCR	0.03	0.21	14.00	No	No
2 KQDKHYGY	0.03	0.00	16.28	No	Yes
3 IEHTLNEK	0.03	0.04	21.93	Yes	No
4 TLNEKRIL	0.03	0.36	15.99	Yes	Yes
5 DELTRHIR	0.04	0.74	11.32	Yes	No
6 PLQHHNLL	0.04	0.35	21.00	No	No
7 KKYRPETD	0.04	0.00	13.49	No	No
8 IQNGDWTF	0.04	0.31	20.21	No	Yes
9 KPMGPLLV	0.04	0.06	28.00	No	Yes
10 KSYLRSLR	0.05	0.59	19.81	Yes	Yes
11 LDLHQTYL	0.05	0.06	20.21	Yes	No
12 AKELVVVY	0.05	0.00	17.97	Yes	No
13 QDDARKLM	0.06	0.24	21.00	Yes	No
14 SCDVKFPI	0.06	0.00	21.00	No	No
15 AFDGETEI	0.06	0.24	21.94	No	No
16 FYSSYVYL	0.06	0.16	27.20	Yes	No
17 NETHSGRK	0.06	0.00	16.90	No	No
18 KNPDNVVG	0.07	0.00	15.96	No	No
19 FSVMNDAS	0.09	0.82	10.66	No	Yes
20 EKPFGTSY	0.10	0.00	14.00	No	No
21 NFLEVGEY	0.11	0.19	27.84	Yes	Yes
22 CNGGHWIA	0.12	0.04	21.77	No	No
23 RGERRGAP	0.12	0.14	28.00	No	Yes
24 KQAHPDLK	0.14	0.02	14.00	No	No
25 RIILDRHR	0.16	0.80	21.00	No	No
26 YASLRSLV	0.17	0.91	11.10	Yes	Yes
27 PRDANTSH	0.21	0.15	21.00	No	Yes
28 AANRSHMP	0.23	0.13	18.24	No	No
29 FHMYFMLR	0.23	0.86	20.10	Yes	Yes
30 AKGVETAD	0.28	0.03	20.19	Yes	No
31 RVLGRDLF	0.46	0.20	21.55	Yes	No
32 AARYKFIE	0.54	0.05	14.00	No	No
33 GQLMALKQ	0.54	0.88	20.66	Yes	Yes
34 HNLIEAFE	0.62	0.75	21.00	No	Yes
35 DYVRSKIA	0.70	0.59	14.76	Yes	No
36 TEVMKRLV	0.78	0.28	14.00	Yes	No
37 QGIIDKLD	0.86	0.44	21.78	Yes	No
38 DEAIDAYI	0.86	0.60	15.14	Yes	Yes
39 RDFEERMN	0.93	0.42	11.59	Yes	Yes
40 RPIARMLS	0.93	0.48	20.34	Yes	No
41 KAAIAQLR	0.93	0.76	17.28	Yes	Yes
42 EKLLESLE	0.93	0.95	21.50	Yes	Yes
43 PAIISAAE	0.93	0.58	17.56	Yes	Yes
44 NAIIQELE	0.93	0.21	30.06	Yes	No
45 AAALDRMR	0.93	0.93	20.54	Yes	Yes
46 GALLDMIQ	0.93	0.77	20.08	No	Yes
47 KRIIDGFK	0.93	0.25	14.00	Yes	No
48 QDMANWVM	0.93	0.51	11.68	Yes	Yes
49 QVFMRIME	0.93	0.00	17.23	Yes	Yes
50 VQTLAAYE	0.93	0.46	14.59	Yes	Yes
51 EEMVSKLK	0.93	0.75	17.40	No	Yes
52 EKLEATIN	0.93	0.63	13.70	Yes	Yes
53 EQMQREIF	0.93	0.48	15.91	Yes	No
54 ESMAERFA	0.93	0.60	21.00	Yes	Yes
55 KELQRIFW	0.93	0.60	13.05	Yes	Yes
56 NDFEDMMT	0.93	0.91	17.41	Yes	Yes
57 RAFQELLE	0.93	0.82	22.00	Yes	Yes
58 RSFDDAMA	0.93	0.58	14.00	No	No
59 SRTRELLA	0.93	0.63	21.00	Yes	No
60 ADFKAAVA	0.93	0.86	21.00	Yes	Yes
61 EDLVERLK	0.93	0.00	17.52	No	Yes
62 KKLQKLID	0.93	0.01	26.41	Yes	No
63 QTLAQLSV	0.93	0.83	13.50	Yes	No
64 ADFKAQFT	0.93	0.83	10.40	Yes	Yes
Total simulation time = 1175.8					

[†]Columns are as follows: (1) Amino acid sequence. (2) Confidence: I-sites score for motif 8055. (3) Helicity: fraction helix observed in the MD simulation. (4) Length of the simulation. (5) Yes if the sequence is a helix in the parent structure; otherwise, no. (6) Yes if the peptide had acetyl and methyl amide capping groups; otherwise, no.

Measuring Helicity in MD Trajectories

A configuration was considered to be alpha-helical if its backbone atoms (N, CA, C, and O) deviated from an ideal helix by <1.5 Å root-mean-square distance (RMSD).

The total helical content (*helicity*) of the trajectory was the number of alpha-helical configurations divided by the total number of configurations.

Total helicity (H) for a set of peptides k was calculated by summing the total time spent in the helix conformation, as follows:

$$H = \frac{\sum_k h(k)*t(k)}{\sum_k t(k)} \quad (4)$$

where h(k) is the *helicity* of peptide *k, t(k)* is the length of the trajectory.

AGADIR predictions

AGADIR[41] predictions of helicity were calculated using the March 2001 version of the web-server (http://www.embl-heidelberg.de/services/), pH 7.0, T = 278 K, ionic strength = 0.1. Both un-capped and N-terminally acetylated, C-terminally amidated sequences were submitted. Logistic regression analysis of the combined AGADIR and I-sites predictions was done using Excel (Microsoft, Redmond, WA) and other programs.

Correlations

Pearson correlations (r) are calculated using the standard equation:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (5)$$

The significance of the correlations can be measured by computing the probability, P, of finding a correlation r, given that the true correlation is zero, and having N observations:

$$P = erfc(|r| \sqrt{N/2}) \quad (6)$$

where *erfc* is the complementary error function.

RESULTS AND DISCUSSION

Table I summarizies the data for the 64 peptides studied. Here we present evidence that the simulations were sufficiently equilibrated to draw broad conclusions over groups of peptides. Then we show that the observed helicities correlate with I-sites *confidence* values as well as with AGADIR helix propensity values. We also show that capping the ends of eight-residue peptides increases *helicity*.

Conformational Equilibrium

The measure of *helicity* in the simulations is accurate to the extent that the conformational space of the peptides is representatively sampled. Although individual peptides may not have been well sampled, the collective sampling of groups of peptides is likely to have been good, and conclusions can be drawn about the collective energy landscape, for example, of all peptides with higher than 0.50 confidence, or all capped peptides.

Molecular dynamics trajectories can be used to characterize the energy landscape of the system.[42] A qualitative assessment of the degree of equilibrium for each peptide was made as follows. The time course of the simulations was plotted using the clustered conformational states, numbered by cluster size with number one being the largest cluster. Figure 3 illustrates clustered trajectories for eight representatives of the 64 simulations. The trajectory was then divided into two equal parts and the first part compared with the second. Some simulations clearly converged, sampling from the same set of conformations in the first half of the trajectory as in the second half (for example, see GALLDMIQ). Other peptides showed rough energy landscapes where trapped conformations persisted for many nanoseconds (see NAIIQELE).

We defined three degrees of roughness (or equivalently, degrees of equilibrium) based on the cluster analysis: (1) "Rough" energy landscapes had different first (largest) clusters in the first and second halves of the simulation. Overall, 29 of the 64 peptides had rough energy landscapes similar to the peptides NAIIQELE, KKLQKLID, and RPIARMLS in Figure 3. (2) "Medium" energy landscapes had the same first cluster in both halves, but different second clusters. Twenty-two of the 64 peptides had medium degrees of roughness, like RAFQELLE and DELTRHIR in Figure 4. (3) "Smooth" energy landscapes had the same first and second clusters in both halves. Thirteen of the 64 had "smooth" landscapes, similar to peptides AAALRPMR and GALLDMIQ.

Example of a Peptide Energy Landscape

Figure 4 shows the trajectory for a typical peptide, RPIARMLS, in greater detail. The boxed stereo images show a sampling of the six largest clusters. The clusters satisfy the two clustering conditions defined in Methods, that no two clusters look similar, and no single cluster looks like two species (only the backbone atoms are considered). In some cases, the difference between the cluster conformations is subtle, involving only one or two backbone angles. The arrows indicate the flow of simulation time. Some transitions between conformational states are one-way, suggesting that a transition was made from a metastable state to a significantly lower free-energy state. In other cases, two clusters interchange frequently, suggesting a low barrier and similar energies. In this example, after about 7 ns, the remaining trajectory can be described as "smooth" (see Fig. 3).

Despite the appearance of the trajectory as smooth, one can never be sure that the equilibrium state has been sampled. It is possible that metastable trapped states may have persisted throughout the length of the simulation in some cases (see discussion of EDLVERLK, below). However, this is increasingly unlikely as the length of the simulation increases, and even less likely to be significant

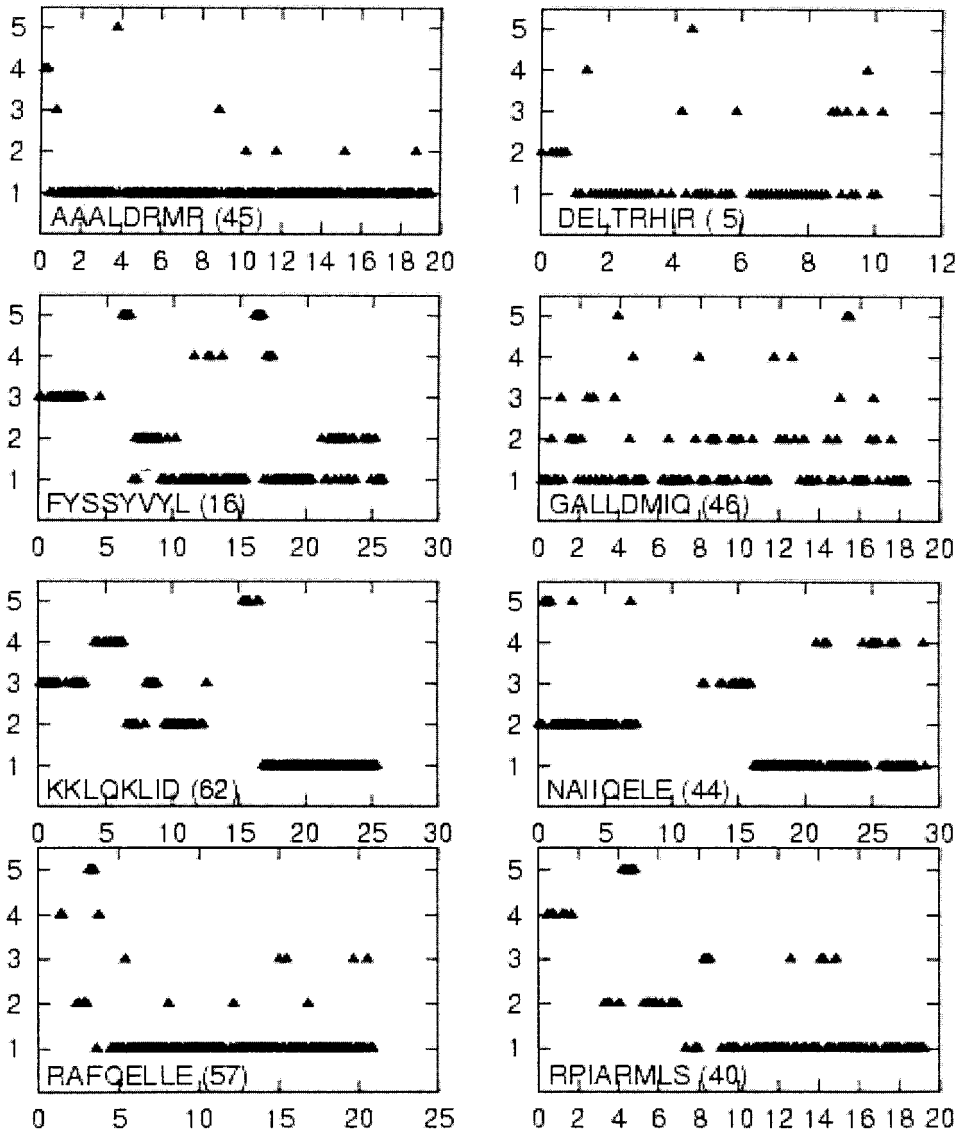

Fig. 3. Sampling of the five largest clusters, plotted versus time (nanoseconds), for selected peptide simulations. The first 1 ns is omitted in all figures and in the cluster analysis. Cluster 1, the largest cluster, is helical for all peptides shown here except FYSSYVYL and KKLQKLID.

when considering collective statistics summed over many peptides.

Confidence Versus Helicity

Table I lists the I-sites confidences and the helicities in MD simulations for 64 peptides. Thirty-three of these peptides had confidences >0.50 and 31 had confidences <0.50. For the low confidence set (594 ns total simulation time), the total helicity was $H = 0.23$. For the high-confidence set (582 ns), $H = 0.56$. Unweighted averages of the helicity are similar to the weighted averages, 0.25 for the low-confidence set, 0.56 for the high-confidence set. The probability that the two sets of data are drawn from the same distribution, calculated using the Student's t test,[43] is $P = 1.94\text{E-}05$. Figure 5a shows the boxed 95% confidence regions for each set.

Although significant differences can be found between two classes of peptides (greater than and less than 0.50 confidence), the data are insufficient to find a significant intermediate class. Better sampling of peptides of intermediate confidence might eventually make this possible. But the two-state result is not unexpected given the two-state behavior of cooperative folding units such as helices.

Helicity in Simulations Versus Helicity in Context

The overall helical content of peptide simulations was found to be 30–70% of the value expected based on database statistics. The range depends on the choice of cutoff value for defining a helix. The low end (30%) uses the definition of helix described in the methods (RMSD < 1.5 Å), whereas the high end (70%) also includes all non-helical members of "helical" clusters—clusters for which

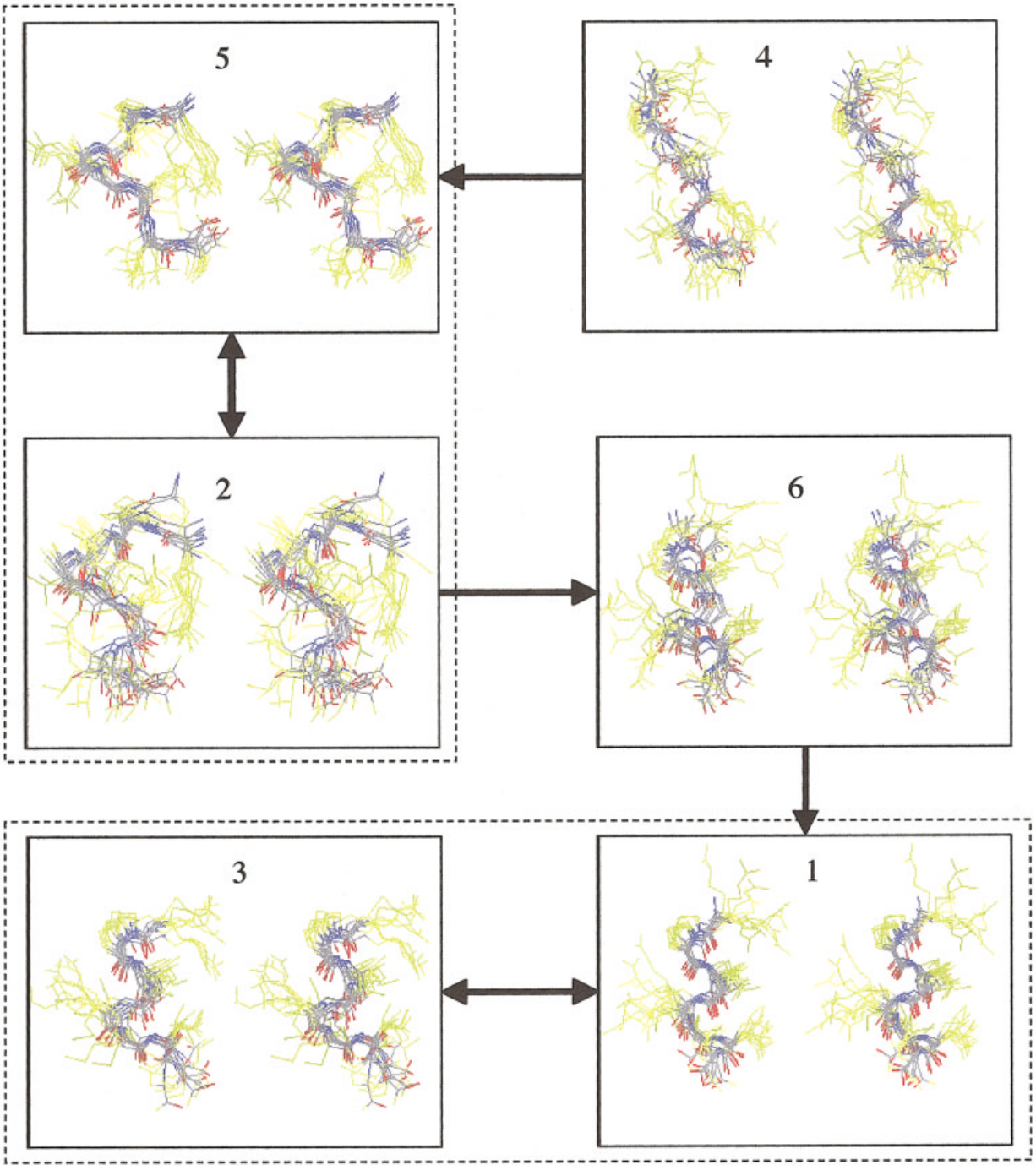

Fig. 4. The rough-energy landscape of peptide 40, RPIARMLS. Each boxed stereographic figure contains a randomly selected sample of the cluster, least-squares superimposed. Side-chain atoms are in yellow. Arrows show the direction of simulation time. Dotted lines surround pairs of clusters that interchange reversibly.

the cluster paradigm is helical in residues 2–7. These "near-helix" conformations might collapse to a helix at lower simulation temperatures. The temperature, 340 K, was chosen to improve sampling. By either measure, the PARM94 force field[37] seems to underestimate helix, contrary to popular belief.[44] Alternatively, the underestimation might be attributable to the absence of the "template effect" in the simulations, as discussed above.

The accuracy of MD as a predictor of helix can be assessed for the peptides studied here. Of the 32 peptides with the lowest *helicity* (average helicity = 0.12), 44% are helix in context. Of the 32 peptides with the highest *helicity* (average helicity = 0.71), 81% are helix in context. If the cutoff for helix prediction is set at 0.16 *helicity* (the optimal setting), then the two-state prediction accuracy is 75% and the false-negatives approximately balance the

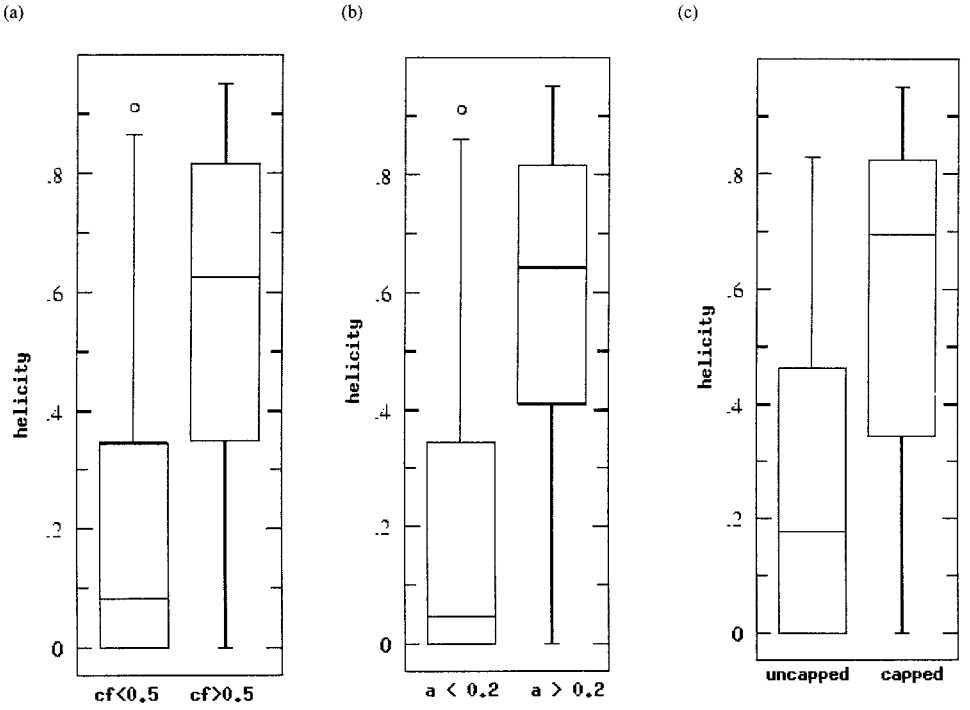

Fig. 5. Statistics for *helicity* of approximately equal-sized sets of peptides. **(a)** Low confidence (<0.50) versus high confidence (>0.50) peptides; boxes mark the 95% confidence intervals. The horizontal line within the box marks the median. The vertical I-beams show the extremes of the data, excluding outliers, circles.[47] **(b)** Helicity for low-propensity (<0.2) and high-propensity (>0.2) peptides accoridng to AGADIR. **(c)** Helicity for uncapped versus capped peptides.

false-positives. This result is comparable to the accuracy of the I-sites method on the 64 peptide set (73%). These numbers cannot be directly compared with the accuracy of secondary structure prediction algorithms[1,2,5,6] which use three single-residue states, whereas here we have two eight-residue states.

If we accept the results of the simulations as the gold standard for helicity of peptides in isolation, then the missing overall helicity, and the missed predictions, must be attributable to the missing tertiary interactions. However, the simulations are imperfect. The presence of a large fraction of near-helix conformations suggests that the underprediction is mostly a temperature effect. Also, many of the missed predictions can be rationalized without invoking the template effect. Five representative cases are discussed in detail further on.

Capping and Helicity

A significant increase in helicity of eight-residue peptides is seen upon capping the ends N-terminally with an acetyl group and C-terminally with a methyl-amide (Fig. 5b). The test for independence (*t* test) of helicity values (*H*) indicates that capped versus un-capped helicity values are unlikely to be chance occurences ($P = 1.0\text{E-}05$).

This result can be partially explained by salt-bridge interactions with the charged ends. Some of the uncapped peptides are observed in trapped non-helical states involving salt-bridges with the charged termini. These conformations would be impossible if the ends were capped (see discussion of NAIIQELE, below). There is no correlation between capping and *confidence*. Therefore, the *confidence*/*helicity* statistics reported above are not the result of capping.

AGADIR and Helicity

AGADIR is a knowledge-based potential function for helix prediction. It is trained on experimental data for designed short peptides (circular dichroism and NMR). Helix propensity predictions using the AGADIR method can be compared to *helicity* in the same way as the I-sites confidence values. The 30 peptides with the lowest propensities (<0.05) had significantly lower ($P = 1.95\text{E-}05$) helicities than the 34 peptides with high propensities (≥0.05). Figure 5c shows the boxed 95% confidence regions for *helicity* of peptides of low and high AGADIR propensity. Interestingly, there is little correlation between the I-sites *confidence* values and the AGADIR helix propensities ($r = 0.33$).

AGADIR predictions were low in general (2.5% over the whole set), compared with the helicity observed in the simulations (21% overall). I-sites predictions are high (52% overall). AGADIR helicity predictions for uncapped sequences had lower values and lower correlations. The low correlation between I-sites and AGADIR propensity measures ($r = 0.33$) is not surprising because the dataset of peptides used to develop the AGADIR potential consisted of mostly non-natural sequences and I-sites uses only natural sequences. An optimized linear combination of the two methods gave a better correlation between a sequence score and *helicity*. A slightly better helix predic-

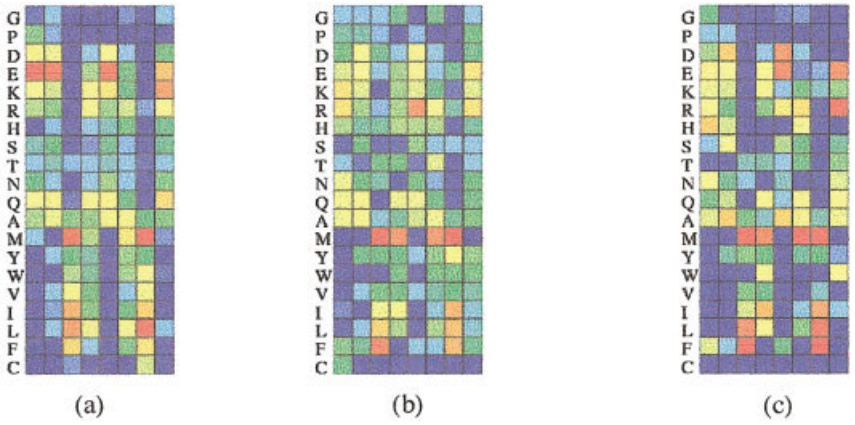

Fig. 6. A survey of eight-residue sequence space explored in this study condensed to sequence profiles by summing the frequencies of each of the 20 amino acids at each position. Position is left-to-right along the bottom, and the color scale is blue (<0.5 times the average amino acid probability), through green (average probability), to red (>2 times the average). Green is the average frequency. **(a)** The I-sites motif 8055 scoring matrix, **(b)** The unweighted amino acid profile of the 64 peptides selected for study. **(c)** The helicity-weighted amino acid profile.

tion score can be obtained using a linear combination, $0.25c + 0.75a$, where c is the I-sites *confidence* and a is the AGADIR propensity. The t test is significantly improved, $P = 4.52\text{E-}07$, indicating that the two prediction methods are probably complementary.

Helix Amino Acid Profile From MD

Figure 6a shows the amino acid probability profile for helices in the context of proteins, summed using the protein structure database as described earlier. We can do a similar summation using the 64 MD simulations. To illustrate the sequence sampling at each position, we simply count the occurences of each amino acid at each position. To get the profile for helix, we weight each count by the *helicity* of that peptide. The results show that certain generalities that we are familiar with in helix sequences also show up in the sequence space of peptide simulations.

Figure 6b shows the profile of the unweighted peptides using a color scale. The overall sampling of amino acids is fairly flat, with a slight bias toward nonpolar side-chains at positions 3 and 7, and polar side-chains at positions 1 and 5, because of the oversampling of higher-confidence sequences (see Methods).

Figure 6c is the profile with *helicity* weighting. The weighted profile resembles the I-sites motif profile (Fig. 6a). Glycines are less probable in all positions except the first. Position 3 prefers hydrophobic side-chains and position 5, polar side-chains. The weighted profile predicts helix in the database much better than the unweighted profile (data not shown), even when the peptides used here were removed from the database.

Rules and Exceptions

The general trend in this study is that peptides that match the amphipathic helix sequence pattern form more helix in MD simulations. However, some of the "outlier" peptides point to problems and possible improvements in the bioinformatics-based method as well as new strategies for peptide simulations. Here we discuss first a typical case, then some of the more obvious false-negatives and false-positives.

A Typical Case
RPIARMLS

This peptide is predicted to be helical (cf = 0.93), and does form a helix after 7 ns of a 20-ns trajectory. Clustering of conformations of this peptide identifies six distinct conformational states that account for 72% of the conformations observed in the trajectory. These six conformational states are shown in Figure 4. One of the trajectory images in Figure 3 shows the order of occurrence.

We find relatively quick back and forth exchanges between clusters 2 and 5 as well as between 1 and 3, indicating lower energy barriers between these states. States 2 and 5 have a salt-bridge interaction between R1 and the (un-capped) C-terminal carboxylate, but lack the I3, M6, L7 hydrophobic cluster. Although the charge repulsion is present between residues R1 and R5 in the helical state, it is minimized by the rotation of R1 psi angle. This leads to sacrifice of one $i \geq i + 4$ backbone H-bond but does not disrupt the helical state.

We also observe (in Fig. 3) that once states 2 and 5 are visited, there are no back transitions to state 4, indicating that state 4 may have considerable higher energy compared with 2 and 5. A similar lack of back transitions is also observed from states 1 and 3 to 2 and 5. A more careful analysis of first passage times and related survival probabilities may provide a complete picture.

False-Negatives: Sequences That Formed Helix but Were Predicted Not to Form Helix
DELTRHIR

In the zinc-finger protein (PDB code IZAA, chain C, residues 20–27), this sequence forms a helix with an internal salt-bridge (D1-R5) and a hydrophobic contact (L3-I7). The simulation reproduces the structure in the context of the protein, but the I-sites confidence is low

because many of the sequence homologs of IZAAC do not conserve a hydrophobic side-chain at position 26 (peptide residue I7); instead, Q and other polar residues align to that position. If all homologs are ignored, then the confidence increases to 0.49. This outlier points to a possible error in assuming that the local structure is conserved in homologs. It also points to an underestimation of the importance of salt-bridging side-chains in the I-sites score.

KSYLRSLR

This sequence comes from the tyrosine kinase domain (PDB code IIRK, residues 1085–1092) where it forms a helix. The 19.8-ns trajectory contained 65% helix. This false-negative I-sites prediction (cf = 0.05) resulted from bad alignment data and the assumption that aligned sequences have the same structure. If the peptide is considered on its own, the helix confidence is much higher (cf = 0.44). Closer inspection of the sequence alignment reveals gaps and insertions in this eight-residue window. The homolog sequences cannot have a helix in this same position, because gaps or insertions would disrupt the helix. A suggested improvement would be to trim sequences from the alignment wherever gaps or insertion occur, and only consider the gapless alignments. This might amplify the true conservation pattern and improve the accuracy of the *confidence* estimations.

False-Positives: Sequences That Were Predicted to Form Helix But Did Not

QVFMRIME

This peptide is predicted to form helix (cf = 0.93), and does so in context (2DLD, 159A–166A), but forms <1% helix in the simulation. The largest cluster is helix on both ends, but the central methionine (M4) has a positive phi angle, which is energetically unfavorable. The AMBER parm94 forcefield assigns an energetic penalty to positive phi angles for non-glycine residues, but it is not known whether the energy balance is correct.

EDLVERLK

Residues 2–8 of this peptide are helical in the context of the protein (PDB code 1QOR, residues 195A–202A), but residue 1 is not. The 17.5-ns-long MD simulation, however, shows no helix formation at all. Like charges occurring at positions 1 and 5 would in principle repel each other in a helix, but in the native conformation they do not, because the side-chain of E1 can rotate away from E5. However, positions 2,6 and 5,8 are oppositely charged pairs, forming salt-bridges that would stabilize a helical state. Positions 3,4 and 7 would fall on one side of helix forming a hydrophobic cluster. Despite the circumstantial evidence of intrinsic helicity, the simulation converged on a non-helical conformation in which the two native salt-bridges form (D2-R6, and E5-K8) but the hydrophobic cluster does not. By eyeball inspection, this state seems like it would have a higher free energy than the native state, which has the hydrophobic cluster in addition to the two salt-bridges. If so, then the simulation after 17.5 ns is still trapped in a local minimum. Several other high-*confidence* peptides are also found in putative "trapped" states (NAIIQELE and RPIARMLS), which may be identified as long-lived, non-native conformations stabilized by favorable interactions, usually salt-bridges.

NAIIQELE

This peptide is predicted to form helix (cf = 0.93), and forms helix in the context of the protein (1BGW 1047–1054). But the peptide forms helix in the simulation only after about 16 ns of a 30-ns trajectory. In the first 16 ns, it passes through two trapped states (clusters 2 and 3 in Fig. 3), both involving a salt-bridge between the un-capped N and C termini. The C-terminal glutamate provides two carboxylate groups to pair with the free N-terminus, increasing the chances of this interaction. These intermediates do not form a hydrophobic core. A sharp transition occurs at 13 ns with the flipping of a single dihedral. A 180° flip of the Q5 backbone psi angle allows two helical H-bonds to form and brings together the I4 and L7 side-chains to form a native contact that persists for the remainder of the simulation. However, the peptide persists in a second trapped state resembling a broken helix for another 3 ns. This is attributable to the continued formation of a salt-bridge between the termini. The presence of trapped states over several nanoseconds simulation time indicates relatively high barriers (~several kT) between these states and the helical state. A better sampling of such energy landscapes will require new approaches, such as the replica exchange method.[45]

Potential Improvements in I-sites and Other Knowledge-Based Scoring Functions

To better match the results of simulations and experiments, the statistical score should approximate a free energy. Missing in the current scoring function is the pairwise covariance between positions in the sequence. A covariance-based score would capture not only charge–charge interactions but also the general non-additivity of probabilties in pairwise and multibody interactions. Even two-body interactions are not additive by nature. For example, it is obvious that the loss of one of the two side-chains forming a hydrophobic contact does not just cut the interaction term by half.

The presence of conformational traps, even for a short peptide, indicates a need for better simulation strategies, such as the replica exchange method,[45] that allow much more efficient sampling of a rough-energy landscape. Implicit solvent models could also speed the sampling of the peptide conformational spaces.

CONCLUSIONS

Meaningful correlations between conformational space and sequence space of short peptides have been found by sampling both spaces, but the correlations are broad in nature because of the sparcity of sampling in both spaces. Amphipathic helices fold primarily via intrinsic, local interactions, not through the template effect of its context in the protein. The image of folding painted by Anfinsen[46]—helices and other local structure motifs "flickering" in and out as cooperative tertiary contacts lock them into place—is supported by our findings.

REFERENCES

1. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19: 55–72.
2. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
3. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
4. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2002;301:173–190.
5. Asai K, Hayamizu S, Handa K. Prediction of protein secondary structure by the hidden Markov model. Comput Appl Biosci 1993;9:141–146.
6. Salamov AA, Solovyev VV. Protein secondary structure prediction using local alignments. J Mol Biol 1997;268:31–36.
7. Parry DA. Coiled-coils in alpha-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. Biosci Rep 1982;2:1017–1024.
8. Waltho JP, Feher VA, Merutka G, Dyson HJ, Wright PE. Peptide models of protein folding initiation sites. I. Secondary structure formation by peptides corresponding to the G- and H-helices of myoglobin. Biochemistry 1993;32:6337–6347.
9. Viguera AR, Jimenez MA, Rico M, Serrano L. Conformational analysis of peptides corresponding to beta-hairpins and a beta-sheet that represent the entire sequence of the alpha-spectrin SH3 domain. J Mol Biol 1996;255:507–521.
10. Ilyina E, Milius R, Mayo KH. Synthetic peptides probe folding initiation sites in platelet factor 4: stable chain reversal found within the hydrophobic sequence LIATLKNGRKISL. Biochemistry 1994;33:13436–13444.
11. De Prat Gay G, Ruiz-Sanz J, Neira JL, Itzhaki LS, Fersht AR. Folding of a nascent polypeptide chain *in vitro*: cooperative of structure in a protein module. Proc Natl Acad Sci USA 1995;92: 3683–3686.
12. Yi Q, Bystroff C, Rajagopal P, Klevit RE, Baker D. Prediction and structural characterization of an independently folding substructure in the src SH3 domain. J Mol Biol 1998;283:293–300.
13. Bystroff C, Baker D. Blind predictions of local protein structure in CASP2 targets using the I-sites library. Proteins 1997;(Suppl 1):167–171.
14. Pearlman DA, Case DA, Caldwell JW, et al. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comput Phys Commun 1995;91:1–41.
15. Weiner PK, Kollman PA. AMBER: assisted model building with energy refinement—a general program for modeling molecules and their interactions. J Comput Chem 1981;2:287–303.
16. Burgi R, Daura X, Mark A, et al. Folding study of an Aib-rich peptide in DMSO by molecular dynamics simulations. J Pept Res 2001;57:107–118.
17. Brooks C, Case D. Simulations of peptide conformational dynamics and thermodynamics. Chem Rev 1993;93:2487–2502.
18. Daggett V. Long timescale simulations. Curr Opin Struct Biol 2000;10:160–164.
19. Hummer G, Garcia AE, Garde S. Helix nucleation kinetics from molecular simulations in explicit solvent. Proteins 2001;42:77–84.
20. Ferrara P, Caflisch A. Folding simulations of a three-stranded antiparallel beta-sheet peptide. Proc Natl Acad Sci USA 2000;97: 10780–10785.
21. Sieber V, Moe GR. Interactions contributing to the formation of a beta-hairpin-like structure in a small peptide. Biochemistry 1996; 35:181–188.
22. de Alba E, Jimenez MA, Rico M, Nieto JL. Conformational investigation of designed short linear peptides able to fold into beta-hairpin structures in aqueous solution. Fold Des 1996;1:133–144.
23. Blanco FJ, Rivas G, Serrano L. A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. Nat Struct Biol 1994;1:584–590.
24. Munoz V, Serrano L. Analysis of i,i+5 and i,i+8 hydrophobic interactions in a helical model peptide bearing the hydrophobic staple motif. Biochemistry 1995;34:15301–15306.
25. Searle MS, Zerella R, Williams DH, Packman LC. Native-like beta-hairpin structure in an isolated fragment from ferredoxin: NMR and CD studies of solvent effects on the N-terminal 20 residues. Protein Eng 1996;9:559–565.
26. Krueger BP, Kollman PA. Molecular dynamics simulations of a highly charged peptide from an SH3 domain: possible sequence-function relationship. Proteins 2001;45:4–15.
27. Moult J. Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol 1997;7:194–199.
28. Koppensteiner WA, Sippl MJ. Knowledge-based potentials: back to the roots. Biochemistry (Mosc) 1998;63:247–252.
29. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4: 187–217.
30. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution [see comments]. Science 1998;282:740–744.
31. Ben-Naim A. Statistical potentials extracted from protein structures: are these meaningful potentials? J Chem Phys 1997;107: 3698–3706.
32. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3;522–524.
33. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. Nucleic Acids Res 2000;28:235–242.
34. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
35. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–4358.
36. Vingron M, Argos P. A fast and sensitive multiple sequence alignment algorithm. Comput Appl Biosci 1989;5:115–121.
37. Cornell WD, Cieplak P, Bayly CI, et al. A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. J Am Chem Soc 1995;117:5179–5197.
38. Darden T, York D, Pedersen L. Particle mesh Ewald: an N. log(N) method for Ewald sums in large systems. J Chem Phys 1993;98: 10089–10092.
39. Ryckaert J, Ciccotti G, Berendsen H. Numerical integration of the Cartesian equations of motion with contraints: molecular dynamics of n-alkanes. J Comput Phys 1977;23:327–341.
40. Berendsen H, Postma J, van Gunsteren W, DiNola J, Haak J. J Chem Phys 1984;81:3684.
41. Lacroix E, Viguera AR, Serrano L. Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. J Mol Biol 1998;284:173–191.
42. Straub JE, Rashkin AB, Thirumalai D. Dynamics in rugged energy landscapes with applications to the S-peptide and ribonuclease-A. J Am Chem Soc 1994;116:2049–2063.
43. Goulden CH. Methods of statistical anaysis, 2nd ed. New York: Wiley & Sons; 1956.
44. Higo J, Ito N, Kuroda M, Ono S, Nakajima N, Nakamura H. Energy landscape of a peptide consisting of alpha-helix, 3(10)-helix, beta-turn, beta-hairpin, and other disordered conformations. Protein Sci 2001;10:1160–1171.
45. Sanbonmatsu KY, Garcia AE. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. Proteins 2002;46:225–234.
46. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
47. Kirkman TW. Statistists to use. 1996. http://www.physics.csbsju.edu/stats/, June 2002.