A Fast Method for Large-Scale *De Novo* Peptide and Miniprotein Structure Prediction

JULIEN MAUPETIT,[1,2] PHILIPPE DERREUMAUX,[2] PIERRE TUFFÉRY[1]
[1]*MTi, INSERM UMR-S973 and RPBS, Université Paris Diderot - Paris 7, 5 rue Marie-Andrée Lagroua Weill-Halle, 75205 Paris, Cedex 13, France*
[2]*Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Institut de Biologie Physico-Chimique and Université Paris Diderot - Paris 7, 13 rue Pierre et Marie Curie, 75005 Paris, France*

Received 27 March 2009; Revised 20 May 2009; Accepted 22 May 2009
DOI 10.1002/jcc.21365
Published online 30 June 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Although peptides have many biological and biomedical implications, an accurate method predicting their equilibrium structural ensembles from amino acid sequences and suitable for large-scale experiments is still missing. We introduce a new approach—*PEP-FOLD*—to the *de novo* prediction of peptides and miniproteins. It first predicts, in the terms of a Hidden Markov Model-derived structural alphabet, a limited number of local conformations at each position of the structure. It then performs their assembly using a greedy procedure driven by a coarse-grained energy score. On a benchmark of 52 peptides with 9–23 amino acids, *PEP-FOLD* generates lowest-energy conformations within 2.8 and 2.3 Å Cα root-mean-square deviation from the full nuclear magnetic resonance structures (NMR) and the NMR rigid cores, respectively, outperforming previous approaches. For 13 miniproteins with 27–49 amino acids, *PEP-FOLD* reaches an accuracy of 3.6 and 4.6 Å Cα root-mean-square deviation for the most-native and lowest-energy conformations, using the nonflexible regions identified by NMR. *PEP-FOLD* simulations are fast—a few minutes only—opening therefore, the door to *in silico* large-scale rational design of new bioactive peptides and miniproteins.

 J Comput Chem 31: 726–738, 2010

Key words: peptide; miniprotein; structural alphabet; structure prediction; coarse-grained force field

Introduction

Although short peptides are known to play many biological functions in cell reproduction, sleep, immune response, regulation, etc. ranging from hormones, neurotransmitters, toxins to antibiotics (e.g., refs. 1–5), the design of peptides aimed at targeting specific molecules is still a challenge for chemical biologists.[6,7]

Currently, our understanding of the relationship between peptide sequences and structures is still limited for three reasons. First, the experimental flow of peptide structure determination based on NMR spectroscopy and X-ray crystallography remains very low. Second, in contrast to proteins, short peptides do not systematically adopt stable well-defined tertiary structures.[8] Third, rapid and accurate *in silico* approaches to engineer peptides for new therapeutic purposes are still missing. Successful application of the Rosetta approach to peptides remains to be determined.[9] All-atom molecular dynamics or Monte Carlo simulations in explicit or implicit solvent models with replica exchanges[10–17] can reach a reasonable structure precision on a small number of peptides and miniproteins, but they are hampered by the computer time and resources. Their performances also remain to be evaluated on a larger set of sequences.[18]

To accelerate conformational search and sampling, one often resorts to simplified representations and energy models, but the main limitation is to preserve the physics of the systems. As a first step, Ichikawa and Dill proposed Geocore,[19] a growing chain algorithm based on a number of discrete ϕ/ψ choices and a sum of hydrophobic and hydrogen-bond interactions. This was followed by PepStr[20] based on secondary structure and β-turn prediction with an energy MD-based refinement, and Peplook[21] based on a Boltzmann-stochastic algorithm coupled to 64 ϕ/ψ backbone combinations. Finally, Nicosai and Stracquadanio developed a generalized pattern search algorithm (GPS)[22] using secondary structure prediction (for peptides with chain lengths >15 amino acids) and an all-atom energy model. Using a benchmark of 42 peptides with 9–20 amino acids, they approached the experimental conformations at 3.2 Å cRMSd (alpha carbon root-mean-square deviation).[22]

In this work, we assess the relevance of a *de novo* approach to predict 3D peptide structures from sequences. *PEP-FOLD* is based on the concept of structural alphabet (SA).[23] We use a Hidden Markov Model (HMM) -derived SA of 27 letters, describing proteins as series of overlapping fragments of four amino acids.[24] This HMM-derived SA description differs, however, from others,[9] in that the conformations of consecutive fragments are not independent, but defined by a Markov transition matrix. In a previous report, we

Correspondence to: P. Tufféry; e-mail: pierre.tuffery@univ-paris-diderot.fr

demonstrated that the application of our 27-state structural alphabet coupled to a greedy algorithm, where the chain is grown one fragment after another, reproduced 20 protein structures of 50–164 amino acids with knowledge of secondary structures and tertiary native contacts.[25]

Here, we escape from any secondary or tertiary structure information and generalize our approach to any natural polypeptide chain. By using support vector machines (SVM) and the forward-backward (FB) algorithm, we are able to predict from the amino acid sequence a limited set of SA letters that encodes the experimental conformation. The assembly of the fragments is then performed using an enhanced version of our stochastic greedy algorithm[25,26] driven by sOPEP, a variant of the OPEP coarse-grained force field.[27]

Materials and Methods

Peptide Test Sets

In this study, we consider two peptide sets of known NMR structures. For performance comparison, we use the *PepStr* set[20] including 42 linear, bioactive peptides with 9–20 amino acids. These monomeric peptides, free of any disulfide bridge, have been characterized by NMR spectroscopy in both aqueous and nonaqueous solutions. For further validation, we have also collected 23 supplementary PDB[28] structures of 10–50 amino acids solved by NMR, and not by X-ray to avoid the impact of crystal packing forces. These structures are also defined as monomers in aqueous solution, are free of any disulfide bond, and contain only natural amino acids. This set, called *PepFold* (see Table 1), contains 10 small peptides with 10–23 amino acids and 13 miniproteins with 27–49 amino acids. For comparison with other techniques, *PepFold* includes the β-hairpin 2gb1F, the three-helix bundle 1bddF, and the three-stranded β-sheet 1e0lF which correspond to the fragments 41–56, 10–55, and 6–34 of 2gb1, 1bdd, and 1e0l PDB entries, respectively.[11–15,18,30,31] Overall, we examine 52 peptides of 9–23 amino acids and 13 miniproteins of 27–49 amino acids.

Table 1. *PepFold* Set. PDB, Protein Data Bank identifier; L, peptide length; α (resp. β): α helix (resp. β-strand) content (in % of the sequence) using STRIDE[29] to assign secondary structures; *M*, the number of NMR models.

No.	PDB	*L*	% α	% β	*M*	RC
Short peptides						
1	1dep	15	0.80		1	–
2	1k43	14		0.43	10	2–14
3	1le1	12		0.67	20	–
4	1le3	16		0.62	20	–
5	1pei	22	0.86		10	–
6	1uao	10		0.40	18	–
7	1wbr	17			32	4–11
8	1wz4	23	0.26		20	1–22
9	2evq	12		0.50	43	–
10	2gb1F	16		0.62	1	–
Mini proteins						
1	1abz	38	0.82		23	1–17, 21–36
2	1bal	37	0.38		56	4–37
3	1bddF	46	0.78		1	–
4	1e0lF	29		0.52	10	–
5	1e0n	27		0.37	10	1–25
6	1f4i	45	0.58		21	1–12, 14–42
7	1fsd	28	0.36		41	2–25
8	1i6c	39		0.33	10	2–11, 16–33
9	1kjk	49	0.35	0.29	25	1–46
10	1psv	28	0.36		32	1–26
11	1ru5	36	0.44		20	5–7, 15–32
12	1vii	36	0.58		1	–
13	2p81	44	0.61		25	11–21, 24–27, 29–31, 33–34, 37.

RC corresponds to the rigid core region, we use here the PDB amino acid numberings. 2gb1F, 1bddF, and 1e0lF correspond to the fragments 41–56, 10–55, and 6–34 of 2gb1, 1bdd, and 1e0l PDB entries, respectively.

SA Letter Encoding from Structure and SA Letter Prediction from Sequence

To encode protein structures as a series of SA letters, we use the forward-backward algorithm (FB) rather than the *Viterbi* algorithm—which calculates the optimal series of SA letter in terms of likelihood—because we prefer to select the *n* most probable SA letters at each position and not only the optimal one. Protein reconstructions are performed using 3D fuzzy trajectories selecting at each position the SA letters with a probability higher than 10^{-6}, leading on average to 4.0 letters by position.

Predicting SA letters of peptides from sequences is challenging because the natural peptide flexibility induces fuzziness in the sequence-structure relationship, and there is not a sufficient number of peptide structures available to learn SA predictors. Here, we use SA predictors learnt from proteins, and apply them to peptides. Our nonredundant (30%) protein set includes 3672 chains from the PDB, solved by X-ray diffraction, with more than 30 amino acids. It corresponds to 5114 continuous fragments (SA Markovian chain breaks on missing residues) and ~860,000 4-residue fragments. This set was divided into a learning set (182,000 fragments) and a validation set (remaining data), and we verified that learning set does not contain any protein entry related to our *PepStr* and *PepFold* peptides.

To predict SA letters from the amino acid sequence, we use the following two-step procedure. The first step predicts the SA letter profiles at each independent position, using an amino acid profile obtained by running PSI-BLAST[32] against the UniRef data-bank.[33] This profile is used as input of a SVM. We use the amino acid profile of each 4-residue fragment enlarged by two residues on both sides. By using a 8-residue window, we have a 20×8 dimensions vector as SVM input to predict one SA letter. The SVM output is a profile of L-3 $\times$ 27 probabilities, which gives the predicted probability that each SA letter fits each fragment of the protein. In a second step, these probabilities are used as input data in the FB algorithm along with the HMM-SA model. Overall, our learning procedure and databases (UniRef and nonredundant PDB set) for deriving SA letters from sequence are free of any interference with our proteins.

We use the SVM facility that comes with the e1071 package of the statistical analysis program R (http://www.R-project.org). This package is built on top of the libSVM library developed by Chang and Lin (http://www.csie.ntu.edu.tw/~cjlin/libsvm). Multiclass classification if achieved using the one-versus-one method, a

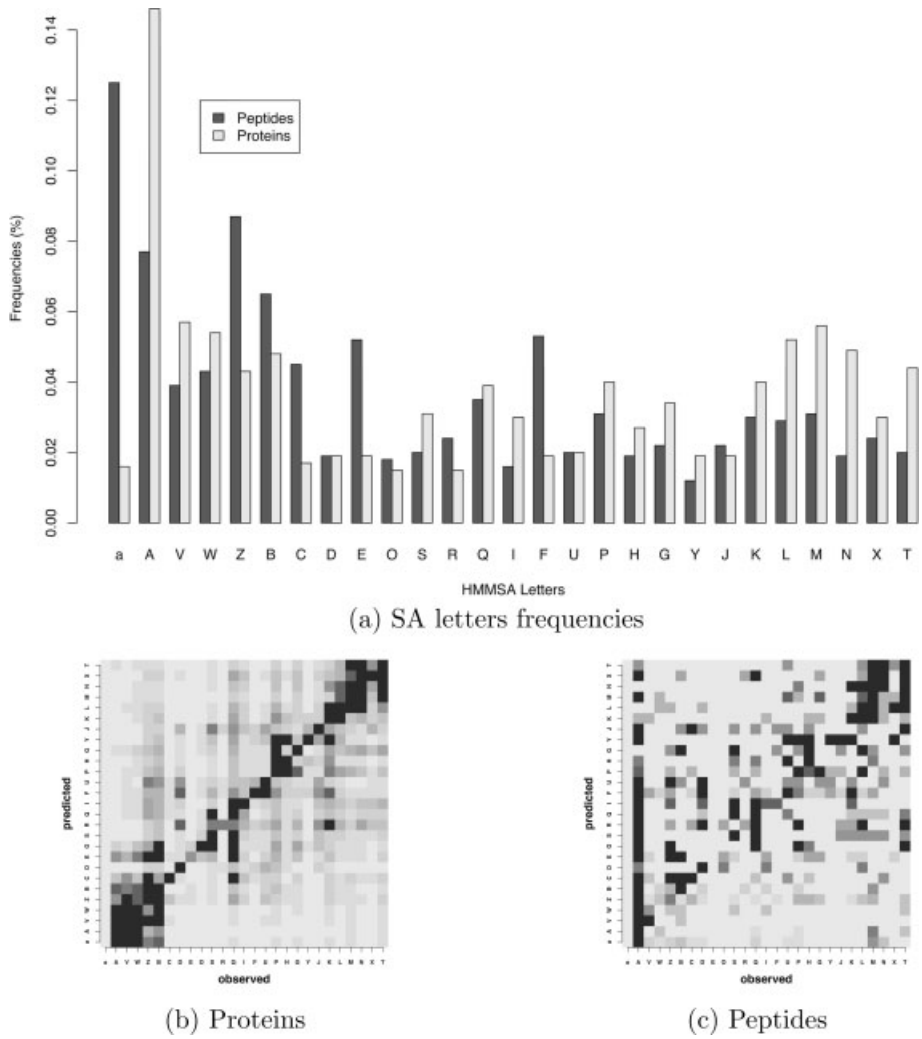

Figure 1. (a) Observed frequencies of SA letters in proteins and peptides. HMM-SA letter frequencies are computed for two encoded sets of structures: our peptide sets and a nonredundant (30%) protein set of 1055 PDB entries. (b, c) Structural alphabet first rank exact prediction (Q^1_{27}) per SA letter. Darker cells correspond to higher occurrences. SA letters are sorted by fragment stretch (A being helical, T being the most extended). Ideal predictions are located in the diagonal. (b) Validation results on proteins. (c) Results on peptides.

strategy that is usually considered having the best performance. We used a radial basis function kernel, i.e., the default one. In this study, we kept the error cost to 1, and used the gamma value internally calculated by default in the package (1/(data dimension)). Finally, since the different SA letters have uneven occurrence frequencies (see Fig. 1a), each SA letter is associated a weight in order to compensate for these differences.

3D Structure from SA Profiles and Greedy

To build 3D structures from a SA profile, we use an enhanced version of our greedy algorithm,[25,26] where instead of applying the forward and backward incremental operators, we develop a zip operator to start the building process at any position of the structure, alternatively adding residues at each side of the growing structure (Fig. 2A). To this end, two new parameters are introduced : a starting position i and a step S, i.e., the number of growing steps one side before switching to the opposite extremity. After extensive tests, we found that for 9–50 residue systems, a S value of 1 is appropriate and the starting position can be selected randomly.

Main Chain Positions

The all-atom backbone coordinates are derived from the HMM-SA prototypes. The transformation matrix used in the greedy algorithm to superpose the α-carbons is also used for other heavy atoms of the backbone (N_i, C'_i, and O_i) and the C'_{i-1}, O_{i-1} when rebuilding from N to C terminus. This operation preserves at best the peptide bond geometry, except that our mean and standard deviation of CA_{i_1}–C'_{i-1} bond lengths is 1.52 and 0.04 Å vs. 1.52 and 0.01 Å (from PDB

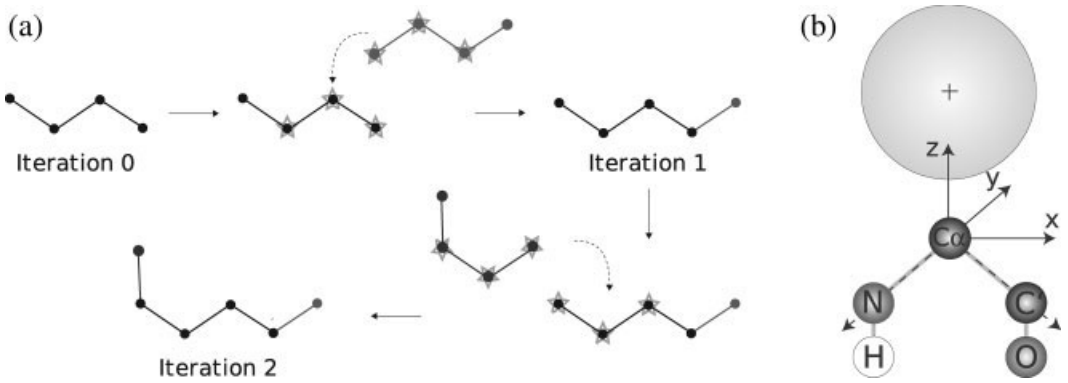

Figure 2. (a) Greedy zip operator, (b) sOPEP coarse-grained model.

analysis). This marginal deviation has no impact on the formation of H-bonds, however, since we generate the hydrogen atom coordinates in a local repair defined by the C'_{i-1}, N_i, and CA_i atoms (see ref. 27). Note that the nitrogen of the N-terminal residue has no hydrogen bound.

Side Chain Positions

Following previous studies,[27] the side-chain beads are positioned using precalculated centroid values depending on the N_i, CA_i, and C'_i coordinates (see Fig. 2B).

Coarse-Grained Potential

We start with the optimized version 3.1 of the generic OPEP force field[27] used for protein folding[34] and aggregation.[35,36] However, OPEP is designed for molecular dynamics in cartesian coordinate space whereas the greedy algorithm works in a discrete space and uses well-defined fragments of four-residues. In addition, OPEP uses the positions of the backbone N, H, Cα, C, and O atoms and one bead for each side-chain (Fig. 2B), while our early greedy algorithm uses a Cα trace. To address both aspects, we design (i) a procedure to generate the positions of all OPEP particles and (ii) sOPEP, a simplified OPEP version adapted to a greedy algorithm.

sOPEP Formulation

The OPEP function is expressed as a function of local terms (bond lengths, bond angles, improper torsions of the side-chains and the peptide bonds, and torsions), nonbonded terms, and H-bond terms.[27] The H-bond potential consists of two-body and four-body terms. Overall OPEP version 3 consists of 261 weighted energy terms.

Because greedy assembles well-defined fragments, our local terms in the present exercise limit to the E_ϕ torsional backbone term expressed by a quadratic polynomial [see eq. (3) in ref. 27]. We retain the OPEP analytic forms of the hydrogen-bonding potential and the nonbonded interactions between the main-chain atoms and between the main-chain and side-chain atoms described by eqs. (6)–(12) in ref. 27. We found, however, necessary to refine the OPEP side-chain side-chain potential term because the assembly process can lead to steric clashes.

The original OPEP potential of mean force between two side chain beads i and j separated by r_{ij} was expressed by:

$$E_{\mathrm{SC,SC}}(r_{ij}) = \begin{cases} -\epsilon_{ij} \times C(r_{ij})^6, \text{ for } \epsilon_{ij} < 0, \\ \epsilon_{ij}(C(r_{ij})^{12} - 2 \times C(r_{ij})^6) \text{ else.} \end{cases} \quad (1)$$

with $C(r_{ij}) = \frac{r^0_{ij}}{r_{ij}}$, r^0_{ij} is the optimal interaction distance, and ϵ_{ij} is the energy well depth at the minimum.

Our new formulation controls r^0_{ij} and ϵ_{ij}, but also the minimal distance R^0_{ij} at which the energy starts to be repulsive.

For $\epsilon_{ij} > 0$, we introduce a new parameter p_{ij}.

$$C(r_{ij}) = \frac{r^0_{ij} - p_{ij}}{r_{ij} - p_{ij}} \quad (2)$$

The value of p_{ij} can be obtained for :

$$E\left(R^0_{ij}\right) = 0 \quad (3)$$

This leads to :

$$p_{ij} = \frac{r^0_{ij} - \sqrt[6]{2} \times R^0_{ij}}{1 - \sqrt[6]{2}} \quad (4)$$

On the other hand, for $\epsilon_{ij} < 0$, we use directly R^0_{ij} to control $E_{\mathrm{SC,SC}}(r_{ij})$, using:

$$C(r_{ij}) = \frac{2 \times R^0_{ij} - r^0_{ij}}{r_{ij}} \quad (5)$$

We derive r^0_{ij} and R^0_{ij} values from a nonredundant PDB (chains with more than 30% of sequence identity are discarded[27]). After testing different values for R^0_{ij}, we chose values that fit to the 0.1 quantile of the distributions for $\epsilon_{ij} > 0$ and 0.2 for $\epsilon_{ij} < 0$. ϵ_{ij} values are taken from OPEP version 3.

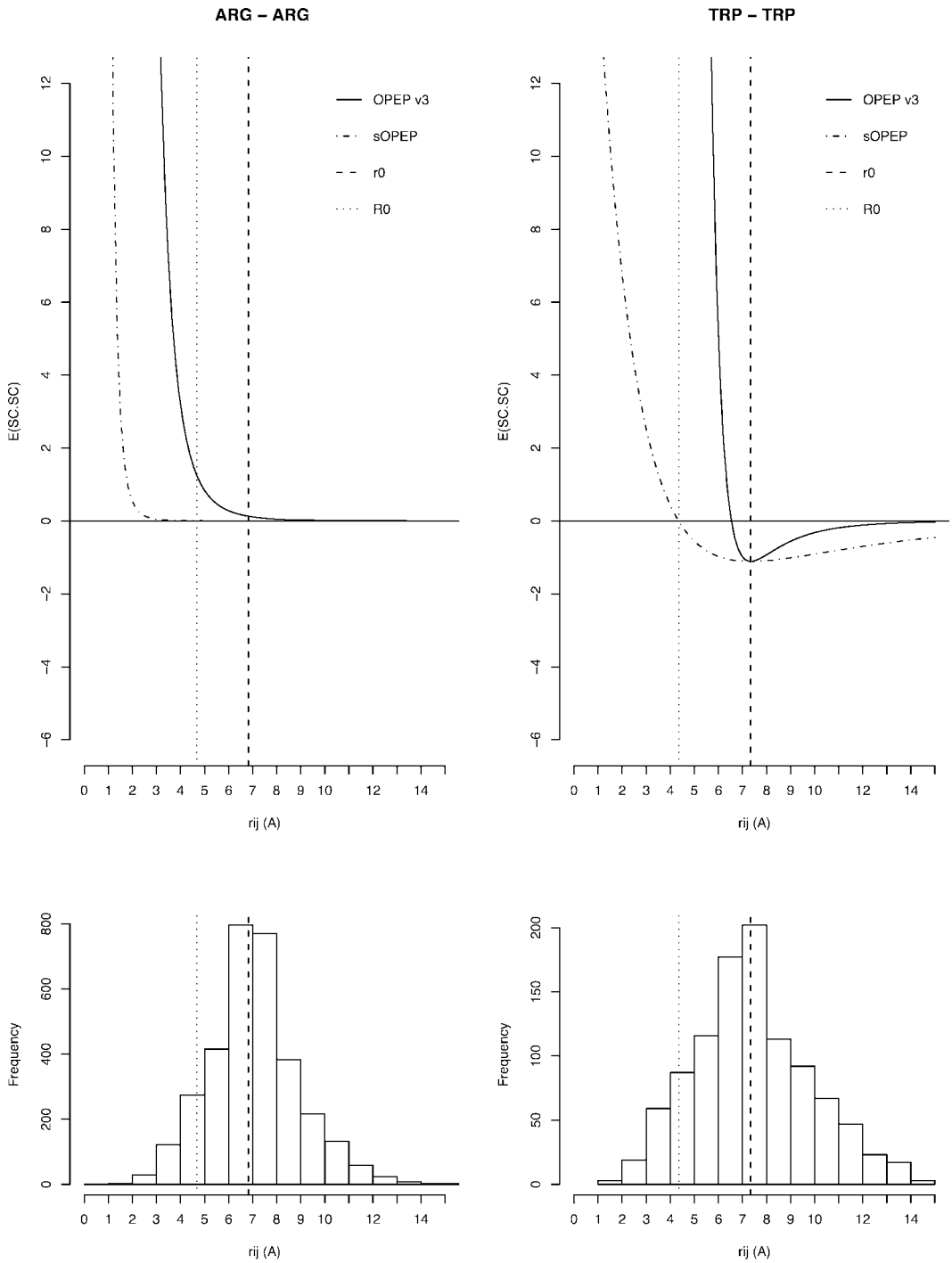

Figure 3. Side-chain interactions with the sOPEP potential. For two different types of side-chain interactions, we present the sOPEP potential (dashed line) and the OPEP formulation (solid line). Discontinuting vertical lines corresponds to *R*0 (dotted line) and r_0 (dashed line).

Figure 3 shows for the Arg-Arg and Trp-Trp side-chain interactions, the difference between the OPEP and sOPEP curves along with the distribution of interacting side chain beads. We clearly see that sOPEP is smoother at short distances, reducing therefore the number of steric clashes during reconstruction.

sOPEP Weight Optimization

The weights of the sOPEP energy components are obtained using our genetic algorithm.[27] Starting from the optimal vector of OPEP version 3.1, we impose that only the 210 weights associated with the

side-chain side-chain interactions are reoptimized and their values are constrained within [0.7;1.3]. The learning set of proteins is very similar to that used during OPEP optimization (except we exclude betanova) and consists of 1abz, 1dv0, 1e0m, 1orc, 1pgb, 2gb1f, 1qhk, 1shg, 1ss1, 1vii, 2ci2, and 2cro-*fisa*. Note that four systems, including 2gb1F, 1abz, 1e0m, and 1vii, are used for learning the sOPEP potential and for testing *PEP-FOLD* as well. This represents a possible 6% bias in our results. However, as shown in our previous work,[27] the force field is very robust and the correlation coefficient between the OPEP vectors derived from two training and validating protein sets is 0.98. In addition, the predictive power of *PEP-FOLD* remains exactly the same if we exclude these four systems among 65 in our analysis.

Simulation Details

To fully demonstrate *PEP-FOLD* efficiency, we report the results of two series of greedy simulations: a first series where the SA letters are derived from the known structures—this corresponds to 3D fuzzy trajectories—and a second series where the SA letters are predicted from the amino acid sequences—*de novo* trajectories.

For each target, we launch 50 independent runs using the fuzzy 3D trajectories and 50 runs using the predicted SA letter profiles.

Several parameters need to be defined for the greedy algorithm. First, to decrease the complexity of the search, the maximal number of prototypes describing each HMM-SA letter is set to three. Hence, instead of a total of 155 prototypes,[25,26] we use 74 prototypes to describe the 27 letters. Second, the present results are based on one single iteration, since we did not observe any improvement in using multiple iterations. We also use a heap size of 3000 conformations at each position, among which 2000 are selected randomly using a Metropolis criterion at $T = 700$ K. At greedy completion, the best model obtained is refined by a Monte Carlo simulation of 300,000 steps, using a random selection of one prototype at each step and $T = 300$ K in the Metropolis acceptance criterion.

Comparison of the PEP-FOLD *Models with NMR Structures*

To assess the quality of the *PEP-FOLD* models with respect to NMR, the conformational diversity of the 50 models is first characterized by cluster analysis using all amino acids, a complete linkage procedure and a threshold of 2 and 3 Å cRMSd for peptides with 9–23 and 27–49 amino acids, respectively. Clustering is therefore carried out without any experimental knowledge.

Next, we use two critical structural data for comparing the models with the NMR reference model. The first RMSd uses the full structures (FS) and amino acids and therefore FS-cRMSd corresponds to a blind trial like CASP. The second RMSd, denoted by RC-cRMSd, uses the experimental rigid cores (RC), thereby excluding the amino acids identified as flexible experimentally. Since NMR PDB entries can contain several models, the NMR reference state is the model designed as best by the authors of the structure if available, otherwise the first deposited model is selected.

To identify the experimental flexible amino acids, we calculate the Cα Root Mean Square fluctuation (cRMSf) of each amino acid as follows. Firstly, for NMR entries with multiple models, all the models are superposed by the iSuperpose online facility (http:// bioserv.rpbs.jussieu.fr/cgi-bin/iSuperpose). Secondly, for NMR entries with a single model, but with chemical shifts in the BMRB database,[37] we use the RCI[38]-predicted RMSf values.

In practice, the rigid core (RC) is defined by excluding the residues considered as flexible, i.e., displaying cRMSf values > 1.5 Å. For 78% (89%) of the residues, we observe that a cRMSf value > 1.5 Å leads to an order parameter S^2 lower than 0.6 (0.7), as predicted by the RCI server. Overall, the full structures and rigid cores match for 27 sequences among 65, and the rigid cores only exclude the N- and C-terminal amino acids for 18 sequences.

Table 2. SVM Structural Alphabet Predictions.

	Proteins		Peptides			
k	Q_{27}^k	$Q\text{cnf}_{27}^k$	Q_{27}^k	$Q\text{cnf}_{27}^k$	cRMSd^{3D}	cRMSd^{P}
1	0.33	0.83	0.17	0.81	1.7	4.0
2	0.49	0.90	0.27	0.89	0.9	2.6
4	0.67	0.96	0.44	0.94	0.6	1.8
6	0.77	0.98	0.59	0.98	0.6	2.0
8	0.82	0.99	0.69	0.99	0.5	0.8

Prediction accuracy at different ranks has been evaluated for our peptide sets compared to a nonredundant protein test set. Q_{27}^k and $Q\text{cnf}_{27}^k$ columns correspond, respectively, to the probabilities to find the exact or at least a compatible HMM-SA letter at rank k. Greedy runs have been launched starting from predicted (P) or 3D (3D) trajectories at rank k, guided by cRMSd (cRMSd) as an objective function. Average cRMSds (in Å) of the generated models are presented for a small test set (1f4i, 1psv, 1le1, and 1pei).

Results

HMM-SA Letter Prediction

As we use a two-step scheme involving (i) fragment—SA letters—prediction and (ii) fragment assembly, a first aspect to analyze is whether the predicted SA letters contain the conformations to accurately reconstruct the structures.

A first observation comes from the analysis of the SA letter frequencies obtained from the encoding of 3D structures. Figure 1a compares the frequencies of the 27 SA letters observed in protein and peptide structures. We find significant variations between the distributions, and notably for the SA letters [a,A] (α-helix cores), [Z] (α-helix extremities), [E] (3.10 helix or type I turn), and [F], the fuzziest letter[24]). This indicates that peptide conformations are less regular than protein structures, which can certainly impact structural prediction from amino acid sequence.

SA prediction scores from sequences are analyzed in Table 2 and Figures 1b and 1c. Table 2 gives the probability to find the exact SA letter (Q_{27}^k) or at least a compatible letter at rank k ($Q\text{cnf}_{27}^k$). At the first rank, it is striking that prediction performance drops from proteins ($Q_{27}^1 = 33\%$) to peptides ($Q_{27}^1 = 17\%$). This difference comes essentially from the helical letter [a], marginally populated in protein structures (Fig. 1a), but occurring frequently in peptide structures. It is the worse predicted in proteins and peptides as shown in Figures 1b and 1c. However, [a] is mostly predicted as [A], i.e., a similar conformation. Accepting more SA letters (Table 2), the

Table 3. *PepFold* Set Modeling Accuracy.

			3D		*De novo*									
						Lowest energy		Best cluster			MP cluster			
No.	PDB	*L*	*FS*d	*RC*d	No. Cl	*FS*d	*RC*d	Rk	*FS*d	*RC*d	PR	*FS*d	*RC*d	qt_{25}
Short peptides														
1	1dep	15	0.7	–	1	1.7	–	1	**1.7**	–	100	1.7	–	1.7
2	1k43	14	0.9	0.6	5	1.6	1.1	1	1.4	1.0	44	1.4	1.0	1.8
3	1le1	12	0.9	–	4	1.0	–	1	**1.0**	–	72	1.0	–	0.9
4	1le3	16	1.7	–	12	1.3	–	10	1.9	–	28	2.4	–	2.6
5	1pei	22	1.5	–	1	1.5	–	1	**1.5**	–	100	1.5	–	1.5
6	1uao	10	1.8	–	3	2.0	–	2	**2.0**	–	82	3.5	–	3.5
7	1wbr	17	3.5	1.0	1	3.5	1.4	1	3.5	**1.4**	100	3.5	1.4	3.5
8	1wz4	23	4.7	4.3	8	5.7	5.2	8	5.0	4.9	28	6.1	5.8	5.9
9	2evq	12	0.6	–	6	0.9	–	2	0.7	–	68	2.0	–	0.8
10	2gb1F	16	1.0	–	10	2.2	–	2	1.3	–	28	5.3	–	2.3
	Mean		1.7	1.4	5.1	2.1	1.8	2.9	2.0	1.7	65.0	2.8	2.6	2.5
Miniproteins														
1	1abz	38	2.7	2.7	9	3.0	2.8	1	**3.0**	**2.8**	48	3.0	2.8	3.2
2	1bal	37	2.5	1.8	35	5.2	4.7	4	4.1	3.7	10	8.0	7.6	5.3
3	1bddF	46	1.6	–	13	3.5	–	8	2.2	–	22	8.9	–	3.9
4	1e0lF	29	1.8	–	31	4.3	–	1	3.1	–	14	3.1	–	6.7
5	1e0n	27	1.7	1.5	12	6.8	6.4	8	6.5	6.6	34	7.6	7.7	6.9
6	1f4i	45	3.3	2.3	15	6.2	5.6	9	4.4	3.7	20	5.7	5.3	6.1
7	1fsd	28	1.9	1.3	7	7.0	6.3	6	4.0	3.9	74	7.4	6.9	7.3
8	1i6c	39	4.6	3.0	41	7.1	5.4	7	6.3	4.3	6	8.4	8.3	8.0
9	1kjk	49	2.4	2.1	38	4.7	4.4	8	**4.7**	4.4	12	9.4	9.6	8.4
10	1psv	28	3.1	3.0	14	5.1	5.1	9	2.6	2.7	40	5.8	5.9	5.6
11	1ru5	36	4.5	2.5	12	6.4	2.7	6	5.1	2.3	44	5.9	2.6	5.4
12	1vii	36	1.8	–	11	4.9	–	5	4.7	–	38	7.4	–	5.0
13	2p81	44	6.1	1.4	11	7.1	3.4	1	5.5	2.8	38	5.5	2.8	5.7
	Mean		2.9	2.1	19.2	5.5	4.6	5.6	4.3	3.6	30.8	6.6	6.1	6.0

PBD: Protein Data Bank identifier. For each sequence of length *L*, we give the results using the fuzzy (10^{-6}) trajectories (see methods) and the *de novo* SA letter profiles. For the *de novo* prediction, we show the number of clusters (Cl), the lowest energy conformation, the best cluster (lowest cRMSd with respect to native) and the most populated (MP) cluster. Each cluster or conformation is identified by its cRMSd with respect to the NMR reference structure using all amino acids (FS-cRMSd denoted as FSd) or the rigid cores (RC-cRMSd denoted as RCd). For the best cluster, we also give its rank among all clusters, and for the MP cluster its population rate. When the lowest energy conformation and the best cluster match, values are in bold. qt_{25}: 25% quantile cRMSd value.

prediction scores for exact letters remain higher for proteins than for peptides, even at rank eight where $Q_{27}^{8} = 0.69$ in peptides vs. 0.82 in proteins.

If now we make the hypothesis that the exact prediction of SA letters is not necessary to reach a native global conformation, we find that the $Q\text{cnf}_{27}^{k}$ values are very similar between proteins and peptides, independently of the *k*th rank. Both Q_{27}^{8} and $Q\text{cnf}_{27}^{8}$ reach 99% for proteins and peptides, suggesting that the prediction of only the eight best letters contains the native conformation information. To verify this, we have performed two types of 3D reconstructions using a cRMSd criterion to guide the greedy algorithm.[24] Results on four proteins are presented in Table 2 (cRMSd columns). The first construction uses the 3D fuzzy trajectories, while the second uses the *de novo* trajectories. Note that in a previous report, we showed that 16 proteins varying from 56 to 247 amino acids were reconstructed at 0.55 Å using our assembly procedure, structural alphabet, and 3D fuzzy trajectories.[25] We clearly see that, by using a 8 SA letter profile, the structural approximations reached by both trajectories are very similar (0.8 Å vs. 0.5 Å).

Overall, this demonstrates that the experimental configurations are perfectly described by our current predicted SA letter profiles.

sOPEP Force Field Effectiveness

A second aspect to be analyzed before *de novo* prediction is whether our energy score coupled to our progressive assembly approach can reconstruct protein structures with fuzzy 3D trajectories.

With the same ensemble of decoys used for OPEP version 3 optimization,[27] we find that sOPEP can identify a native or native-like structure as the lowest energy conformation for 22 among 29 targets (data not shown). This score is very similar to OPEP performance (23 among 29 targets). This shows, as we also find here, the assumed native configurations have higher energies than the *PEP-FOLD* models for some miniproteins. We report in Table 3

(3D columns) the performance of sOPEP applied to the reconstruction of models from the fuzzy 3D trajectories. Since peptides show experimental flexibility, we report for each target the cRMSd of the most-native predicted conformation using full structure (FS) and rigid core (RC). Table 1 reports for each target the RC regions.

Averaged over the 23 systems of the *PepFold* set, the reconstructions differ from experiment by 1.7 Å for the 10 short peptides, and 2.9 Å for the 14 miniproteins. If the RC criterion is used, these values fall down to 1.4 and 2.1 Å, respectively. This clearly demonstrates that our general procedure using sOPEP produces 3D models near the experimental conformations if the SA trajectories are reasonable.

For the peptides 1i6c, 1ru5, 2p81, and 1wz4, however, the best generated structures display a cRMSd greater than 3.5 Å. For 1i6c, 1ru5, and 2p81, this comes from flexibility, the cores deviating by 3.0, 2.5, and 1.4 Å. For the 23-residue 1wz4 peptide, flexibility does not come into play (4.3 vs. 4.7 Å) and the discrepancy is discussed below.

De Novo *Prediction on the* PepFold *Set*

For each target, we have launched 50 independent runs using the predicted eight SA letter profiles. We first present in Table 3 the number of clusters using the 50 final generated structures and a 2 Å and 3 Å cRMSd cutoff for peptides with 9–23 and 27–49 amino acids, respectively. Next, we report the lowest energy conformation with its cRMSd, the best (or most native) cluster with its centroid cRMSd, and the most populated cluster with its centroid cRMSd. In all cases, we give the FS-cRMSd and RC-cRMSd. Finally, the clusters are ranked according to their population rates ($PR_i = \frac{n_i}{N}$, with n_i the number of models in the cluster i, and N the total number of produced models), and we also give the rank of the best cluster. In what follows, models are discussed using their PDB amino acid numberings. Figure 4 shows the superposition of the predicted models on NMR structures for 16 systems.

Short Peptides

Averaged on the 10 peptides, the number of clusters is 5.1 and ranges from 1 to 12. The most populated clusters represent 65% of the generated conformations and their representative conformations deviate from the NMR structures by only 2.8 Å FS (2.6 Å RC), whereas the best clusters deviate by 2.0 Å FS (1.7 Å RC). The average rank of the best cluster is 2.9, but for 5 targets among 10, the best cluster is the most populated cluster. In addition, we note that for 1le3, the best cluster is ranked 10 with a cRMSd of 1.9 Å whereas the most populated cluster is associated with a cRMSd of 2.4 Å.

Interestingly, the performances based on the lowest energy conformations are very similar to those based on the best clusters, with a mean FS-cRMSd value of 2.1 Å (1.8 Å RC). The mean energy differences between the lowest energy conformation and the most populated cluster (resp. the best cluster) are of 1.3 (resp. 0.6) kcal/mol. For 1wbr and 1wz4, however, the lowest-energy states deviate by 3.5 and 5.7 Å FS-cRMSd, respectively. For 1wbr, the decrease in quality comes from flexibility, the rigid core deviating by 1.4 Å. For 1wz4, the *de novo* structure superposes very well on that generated from the 3D fuzzy trajectory, but the predicted structure differs from NMR by a shift of the first α-helix (3–8 vs. 2–7 in NMR) and of the second helix (Fig. 4a). The NMR structure displays, however, a clash in this region, with the backbone oxygens of Gln7 and Asp11 separated by 2.5 Å.

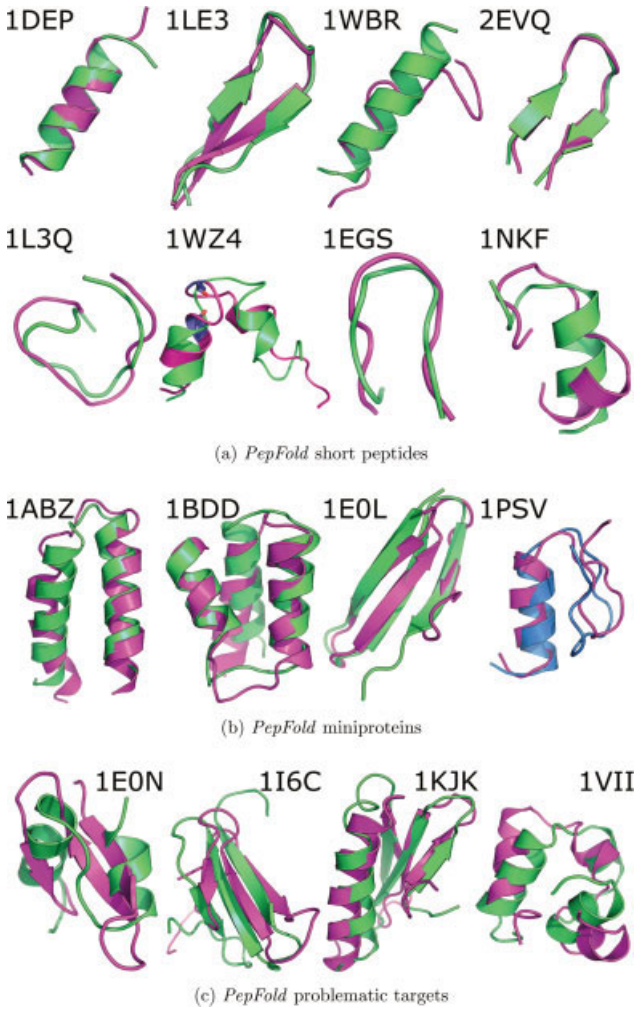

Figure 4. Best predicted and problematic models for *PepFold* set. The reference NMR structure is represented in magenta. It is superimposed either on the lowest-energy model (green) or the most-native, best cluster (blue) for *PepFold* short peptides (a), and *PepFold* miniproteins (b). For problematic targets (c), the centroid of the best cluster is represented (green).

All these results demonstrate that *PEP-FOLD* generates native-like conformations of lowest-energy for the 10 short peptides.

Miniproteins with 27–49 Amino Acids

We now present the results on miniproteins displaying various topologies by NMR. Averaged over the 13 systems, the best cluster deviates from the NMR conformation by 4.3 Å (FS-cRMSd) and by 3.6 Å (RC-RMSd). We observed a mean difference of 2.3 Å between the conformations of the best and most populated clusters. The average lowest energy conformation is at 5.5 Å (FS) and 4.6 Å (RC) cRMSd from the NMR structures, and the energy difference

between the most populated clusters and the lowest energy conformations are on the order of 3 kcal/mol. This indicates that sOPEP is not optimal yet for recognizing near-native from higher RMSd states and the relationship between low energies and cRMSd's can be improved.

Overall, *PEP-FOLD* recognizes the lowest-energy state as the lowest FS-cRMSd structure for the 38-residue 1abz and 49-residue 1kjk targets, and generate near-native conformations (RC-cRMSd < 4.0 Å) for 9 among 13 targets.

For the 39-residue 1i6c, 49-residue 1kjk, and 36-residue 1vii miniproteins, the lowest-energy RC models display native topologies and differ by 5.4, 4.4, and 4.9 Å RMSd from NMR as follows. As seen in Figure 4c, the variation for 1i6c comes from more disordered loops connecting the strands and a shift in the native strand 26–28, modeled as 27–30. For the 1kjk target (PDB residues 11–59), the *PEP-FOLD* model displays three strands at positions 15–18, 23–28, and 33–37 and an α-helix at positions 40–57, i.e., similar to the experimental topology characterized by three strands at positions 16–19, 23–27, and 34–38 and an α-helix 41–57. However, the first strand does not display the native hydrogen bonds with strand 2, and the C-terminal helix has a slightly different orientation. For the 1vii target (PDB residues 41–76), the variation comes from a different orientation of helix H1 (44–49) in the *PEP-FOLD* model.

By contrast, for the 27-residue 1e0n peptide, as illustrated in Figure 4c, the best energy model with a cRMSd of 6.4 Å displays a β-strand packed against two α-helices (9–13 and 23–30) vs. a three-stranded β-sheet by NMR. The SA letter profile at positions 9–13 and 23–30 reveals predicted SA letters with both high α-helix or β-strand characters. While this peptide is the unique system of our set to fold into a non-native topology, we note that PSIPRED also fails to predict the third strand[39] and a series of 10 μs molecular dynamics simulations on a similar WW domain also point to many metastable states with high helical compositions.[40]

Discussion

Recent studies indicate that the *PepStr* algorithm approximated the structure of the 42 *PepStr* set by 4 Å FS-cRMSd.[22] By contrast, the GPS algorithm recently reached an averaged value of 3.2 Å FS-cRMSd on a slight variation of the *PepStr* set.[22] Here, using the exact *PepStr* set of 42 peptides, the best or most-native cluster, the most populated cluster, and the lowest energy conformation are similar, and approximate the native structure by an averaged value of 2.7, 2.8, and 3.0 Å FS-cRMSd and by 2.3, 2.4, and 2.4 Å RC-cRMSd (see Table 4). The energy differences between these three types of structures are marginal, on the order of 0.5 kcal/mol. The predicted models of 19 targets are presented in the Figure 5. For 12 common peptides in aqueous solution, GPS averaged FS-cRMSd is 3.5 Å, whereas *PEP-FOLD* reaches 2.4 Å FS-cRMSd. This improvement in performances of 1.1 Å cRMSd is non-negligible because the GPS results are biased by secondary structure prediction for peptides with chain lengths higher than 15 amino acids.

For eight *PepStr* targets among 42, however, the *PEP-FOLD* lowest-energy conformation differ by more than 4 Å from experiment: 2bta, 1du1, 1e0q, and 1nkf in aqueous solution, and 1g89, 1id6, 1m02, and 2bp4 in nonaqueous solution.

For 2bta, the deviation mainly comes from flexibility: the RC-cRMSd of the lowest-energy state is at 2.4 Å. For its part, 1du1 structure is characterized by one α-helix (Ser2-Ala8) followed by a destructured region in its PDB entry, but it is described as one helix covering 1–15 in the accompanying article.[41] Our *de novo* model is in between, displaying one helix spanning 2–19. The native 1e0q structure is a β-hairpin with two strands at 2–7 and 11–16, while we identify two strands at positions 5–7 and 14–16. Finally, the 1nkf predicted model (resp. native structure) displays a turn region at positions 1–5 (*resp.* 1–4) followed by an α-helix at positions 6–15 (resp. 10–16). However, the experiment uses a saturated concentration of La^{3+} ions,[42] neglected in *PEP-FOLD*.

The three peptides 1g89, 1id6, and 1m02 are fully unstructured by NMR in nonaqueous solution. They are predicted as unstructured (1g89) and essentially helicoidal (1id6 and 1m02), see Figure 5. For its part, 2bp4 displays one long helix experimentally, whereas we find a shorter helix followed by a turn. However, 1g89 structure determination was carried out in dodecylphosphocholine (DPC) and sodium dodecylsulfate (SDS),[43] 1m02 experiment in SDS micelles,[44] 1id6 experiment in dimethyl sulfoxide (DMSO), and 2bp4 in trifluoroethanol (TFE) water mixture. It is clear that the absence of an hydrophobic (SDS) or polar (DMSO, TFE) solvent in our simulations leads to different accessible surface areas.

It is also interesting to compare *PEP-FOLD* with other simulation approaches. To this end, we will use the lowest cRMSd structure, the lowest free energy structure or the lowest effective energy generated by all techniques including *PEP-FOLD*. Our goal here is not to look at the statistical mechanics aspects of protein folding.

Based on Monte Carlo simulations and semiempirical physicochemical potentials, Clarke and Parker calculated with reasonable accuracy the structures of short peptides with 10–20 amino acids, but failed for the 20-residue 1l2y Trp-cage.[13] The lowest energy conformation we report for 1l2y is 2.1 Å cRMSd, i.e., a precision obtained by replica exchange molecular dynamics (REMD) simulations with explicit or implicit solvent models.[18,45,46]

The 36-residue villin headpiece subdomain (1vii) was studied by 1 μs molecular dynamics (MD) in explicit solvent[47] and a 240 ns REMD with a generalized Born solvation model.[48] The all-atom MD trajectory never reached a near-native state below 4.5 Å cRMSd, and the REMD-predicted native state was not associated with the lowest free energy minimum. Here, we reach one native-like cluster with a cRMSd of 4.4 Å ranked fourth according to our population rate.

The 45-residue 1f4i domain was investigated using free energy based all-atom protein folding and worldwide distributed computational resources. By using 2048 processors, a near-native conformation deviating by 3.4 Å cRMSd was located in less than 24 h.[49] Here, we reach one native-like state with a 3.7 Å cRMSd, ranked ninth in our clustering procedure.

The 46-residue fragment of the B-domain of protein A (PDB code 1bddF covering the region 10–55 of 1bdd) was studied by various folding approaches.[30,31] Using all-atom Monte Carlo replica exchange (REMC) simulations, Shakhnovich and Coworkers identified a lowest energy minimum at 6.4 Å cRMSd that was refined to 3.8 Å cRMSd using a second REMC at lower temperature.[31] A similar topology was studied by all-atom discrete molecular dynamics and the engrailed homeodomain protein takes a long simulation time to

Table 4. PepStr Set Modeling Accuracy.

					Lowest energy		Best cluster			MP cluster		
No.	PDB	*L*	RC	No. Cl	*FSd*	*RCd*	Rk	*FSd*	*RCd*	*FSd*	*RCd*	qt_{25}
Aqueous solutions												
1	1a13	14	2−14	1	1.8 (4.5)	1.6	1	**1.8**	1.7	1.8	1.7	1.8
2	1b03A	18	–	12	2.0 (3.8)	–	2	2.0	–	2.5	–	2.5
3	1du1	20	–	1	5.1	–	1	**5.3**	–	5.3	–	5.1
4	1e0q	17	–	3	4.7 (4.3)	–	2	4.5	–	5.0	–	4.8
5	1egs	9	–	2	1.5 (2.3)	–	1	**1.5**	–	1.5	–	1.4
6	1gjf	14	4−14	1	2.5 (4.8)	0.4	1	**2.5**	0.4	2.5	0.4	2.5
7	1in3	12	–	1	2.4 (3.3)	–	1	**2.3**	–	2.3	–	2.4
8	1l2y	20	–	11	2.1	–	6	2.1	–	3.3	–	2.9
9	1l3q	12	–	6	3.3 (4.6)	–	3	3.4	–	3.6	–	3.3
10	1lcx	13	3−12	1	2.8 (2.5)	2.6	1	**2.8**	2.6	2.8	2.6	2.8
11	1niz	14	–	3	2.1 (5.0)	–	1	**1.5**	–	1.5	–	2.0
12	1nkf	16	–	1	4.3	–	1	**4.3**	–	4.3	–	4.3
13	1pef	18	–	1	0.9 (0.6)	–	1	**1.0**	–	1.0	–	0.9
14	1rpv	17	4−16	1	0.6 (2.0)	0.5	1	**0.7**	0.5	0.7	0.5	0.6
15	2bta	15	4−9	1	4.5 (4.6)	2.4	1	**4.5**	2.5	4.5	2.5	4.5
	Mean			3.1	2.7	2.4	1.6	2.7	2.4	2.8	2.5	2.8
Nonaqueous solutions												
1	1c98	10	–	2	3.7	–	1	**3.5**	–	3.5	–	3.7
2	1d6x	13	1−12	2	4.0	3.8	1	**4.0**	3.7	4.0	3.7	4.0
3	1d7n	14	2−13	1	1.0	0.8	1	**1.0**	0.8	1.0	0.8	0.9
4	1d9j	20	10−20	5	5.7	1.5	3	2.2	1.5	3.1	1.4	4.2
5	1d9l	17	4−15	1	1.8	0.9	1	**1.8**	0.8	1.8	0.8	1.8
6	1d9m	18	10−18	1	2.7	1.5	1	**2.5**	1.5	2.5	1.5	2.7
7	1d9o	20	11−20	1	3.3	0.7	1	**3.3**	0.7	3.3	0.7	3.3
8	1d9p	20	10−20	3	6.8	0.8	1	**2.1**	0.8	2.1	0.8	2.1
9	1dn3	15	–	1	1.2	–	1	**1.2**	–	1.2	–	1.2
10	1g89	13	1−12	3	4.7	4.5	1	**4.7**	4.5	4.7	4.5	4.8
11	1hu5	18	2−18	1	1.9	1.7	1	**1.8**	1.6	1.8	1.6	1.9
12	1hu6	18	2−18	1	3.9	3.6	1	**4.0**	3.8	4.0	3.8	3.8
13	1hu7	18	2−18	1	2.2	1.9	1	**2.2**	1.9	2.2	1.9	2.1
14	1id6	15	1−14	1	6.1	5.6	1	**6.1**	5.6	6.1	5.6	6.0
15	1jav	19	–	1	2.0	–	1	**1.9**	–	1.9	–	1.9
16	1kzv	18	2−18	1	2.7	2.4	1	**2.7**	2.4	2.7	2.4	2.7
17	1m02	12	2−11	1	5.1	3.8	1	**5.1**	3.8	5.1	3.8	5.1
18	1myu	12	–	1	3.5	–	1	**3.4**	–	3.4	–	3.5
19	1odp	20	–	1	2.0	–	1	**2.0**	–	2.0	–	2.0
20	1p0j	19	2−19	1	2.0	1.9	1	**2.1**	1.9	2.1	1.9	2.0
21	1p0l	19	2−19	1	2.1	1.9	1	**2.1**	1.9	2.1	1.9	2.1
22	1p0o	19	2−19	1	2.0	1.9	1	**2.0**	1.9	2.0	1.9	2.0
23	1p5k	19	2−18	1	2.0	1.7	1	**2.0**	1.7	2.0	1.7	2.0
24	1qcm	11	2−11	1	2.1	1.6	1	**2.1**	1.6	2.1	1.6	2.1
25	1qfa	13	–	1	0.8	–	1	**0.9**	–	0.9	–	0.8
26	1sol	20	–	1	3.1	–	1	**3.1**	–	3.1	–	3.1
27	2bp4	16	1−15	5	5.6	5.4	2	4.8	4.7	5.4	5.1	5.0
	Mean			1.5	3.1	2.4	1.1	2.8	2.3	2.8	2.3	2.8
All peptides												
	Mean			2.1	3.0	2.4	1.3	2.7	2.3	2.8	2.4	2.8

Starting from SVM predicted profiles, we predict each target with greedy guided by sOPEP objective function. See caption of Table 3 for details. For the lowest-energy prediction, we also report in parentheses the FS-cRMSd predicted by the GPS algorithm.[22]

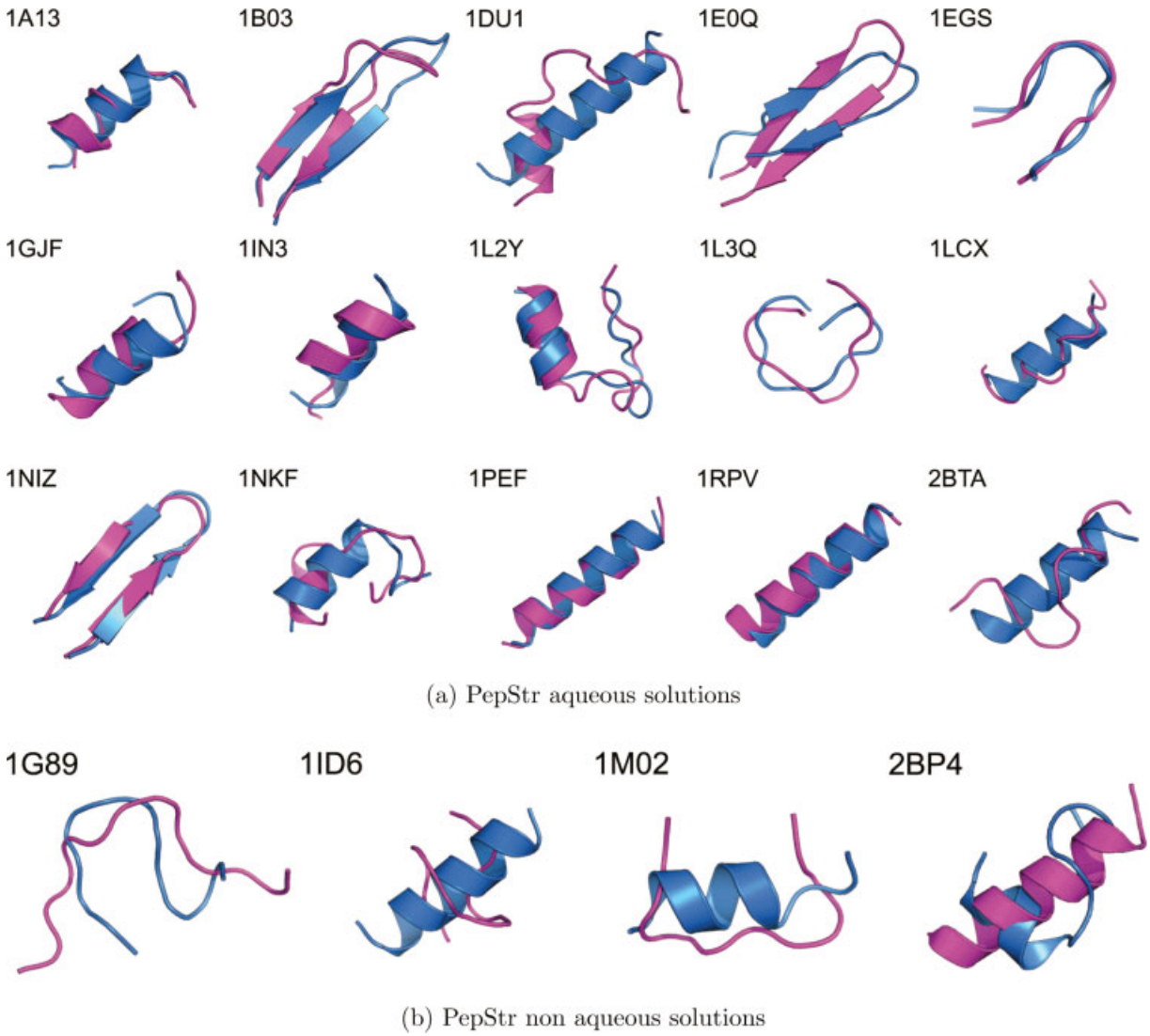

Figure 5. Best predicted models for *PepStr* set. For all targets, the native structure (magenta) is superimposed on the best predicted model (blue) for 15 peptides in aqueous solution set (a) and four problematics targets for peptides in nonaqueous solutions (b). Only the centroids of the best cluster are represented.

reach its near-native state.[50] Here, the most-native *PEP-FOLD* cluster of 1bddF subdomain is at 2.2 Å cRMSd and the lowest-energy conformation at 3.5 Å cRMSd from the NMR structure.

Finally, we briefly discuss the scalability of the method and how its performance varies with the amino acid length. Figure 6A reports the RC-cRMSd of the lowest-energy conformation and of the best cluster (most native) centroid with respect to the NMR structure for all the 38 systems in aqueous solution. The linear regressions show a slight decrease in performance with peptide size. One also note a slight divergence between the two regression lines. As can be seen from the individual dots, the agreement is very good for sizes up to 25 residues, but decreases for longer chains. This indicates that some improvements are required to reach correct behavior for sizes up to 50 residues. Compared to other approaches, *PEP-FOLD* is extremely fast. As illustrated in Figure 6B, execution times vary almost linearly with peptide length. Using a total of 10 CPU Intel Xeon 2.8 GHz, 50 simulations take 30 min and 90 min for 20-residue and 35-residue targets, respectively. This light computational burden allows us to run many simulations and ensure a good sampling of most low energy conformations.

Conclusions

We have presented a new approach for *de novo* structure prediction of peptides from amino acid sequences. *PEP-FOLD* does not rely on any secondary structure information, but rather on the prediction of compatible conformational prototypes of four amino acids along the sequence, and the assembly of these structural alphabet letters using a stochastic greedy algorithm driven by the sOPEP coarse-grained force field.

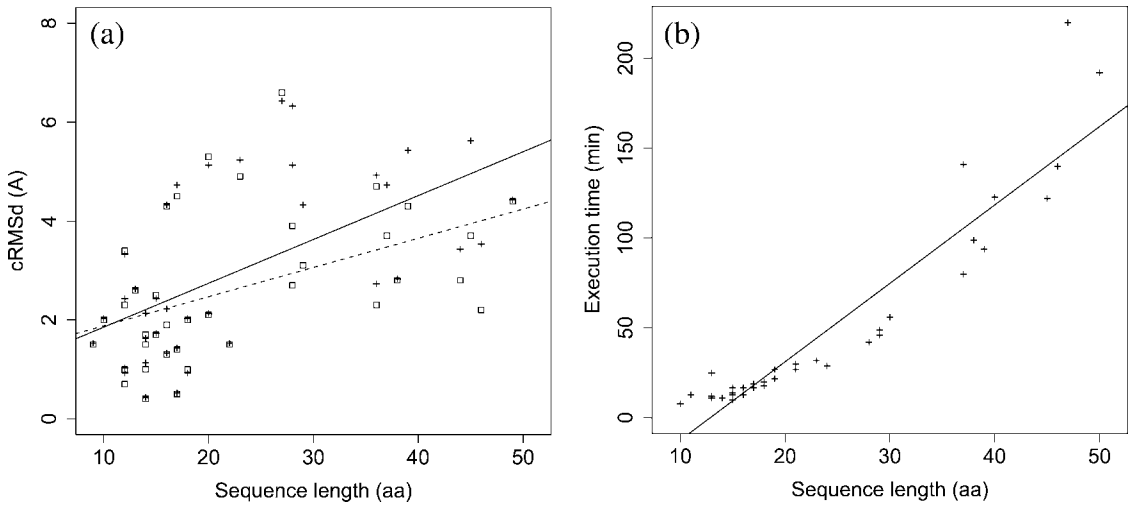

Figure 6. *PEP-FOLD* performance and execution time as a function of peptide length. (a) The lowest-energy conformation (crosses, solid line) and best cluster (squares, dashed line) of all the 38 systems in aqueous solution are reported. Lines correspond to linear regressions. (b) Execution times.

Using a benchmark of 65 sequences, including 52 peptides with 9–23 amino acids and 13 miniproteins with 27–49 amino acids, the predictive power of *PEP-FOLD* can be summarized as follows.

PEP-FOLD generates peptide conformations of lowest-energy approximating the full NMR structures at 2.8 Å cRMSd and the rigid NMR cores at 2.3 Å cRMSd on average. On miniproteins, *PEP-FOLD* finds lowest-energy structure and most-native clusters deviating by 4.6 and 3.6 Å RC-cRMSd from the NMR cores, and by 5.5 and 5.6 Å FS-cRMSd from the full reference NMR structure. The most-native clusters of all miniproteins is always ranked in the top 10 generated models. For all instances, except the 39-residue 1i6c, the predicted topologies are NMR-like.

Although *PEP-FOLD* will get better for miniproteins by improving the accuracy of SA letter prediction and revisiting the sOPEP force field, the speed of the algorithm and the present results open the door to large-scale prediction of peptide structure and peptide engineering. They also provide a strong basis for exploring in a near future linear and cyclic peptides combining both D- and L-amino acids in solution, and peptides in TFE environment or in proximity of membranes.

References

1. Liu, F.; Baggerman, G.; Schoofs, L.; Wets, G. J Proteome Res 2008, 7, 4119.
2. Sang, Y.; Blecha, F. Anim Health Res Rev 2008, 9, 227.
3. Rogge, G.; Jones, D.; Hubert, G. W.; Lin, Y.; Kuhar, M. J. Nat Rev Neurosci 2008, 9, 747.
4. Hong, F.; Ming, L.; Yi, S.; Zhanxia, L.; Yongquan, W.; Chi, L. Peptides 2008, 29, 1062.
5. Erdmann, K.; Cheung, B. W. Y.; Schröder, H. J Nutr Biochem 2008, 19, 643.
6. Zellefrow, C. D.; Griffiths, J. S.; Saha, S.; Hodges, A. M.; Goodman, J. L.; Paulk, J.; Kritzer, J. A.; Schepartz, A. J Am Chem Soc 2006, 128, 16506.
7. Imperiali, B.; Ottesen, J. J. J Pept Res 1999, 54, 177.
8. Gellman, S. H.; Woolfson, D. N. Nat Struct Biol 2002, 9, 408.
9. Bradley, P.; Chivian, D.; Meiler, J.; Misura, K. M. S.; Rohl, C. A.; Schief, W. R.; Wedemeyer, W. J.; Schueler-Furman, O.; Murphy, P.; Schonbrun, J.; Strauss, C. E. M.; Baker, D. Proteins 2003, 53(Suppl 6), 457.
10. Forcellino, F.; Derreumaux, P. Proteins 2001, 45, 159.
11. Jang, S.; Shin, S.; Pak, Y. J Am Chem Soc 2002, 124, 4976.
12. Ulmschneider, J. P.; Jorgensen, W. L. J Am Chem Soc 2004, 126, 1849.
13. Clarke, O. J.; Parker, M. J. J Comput Chem 2008, 29, 1177.
14. Ho, B. K.; Dill, K. A. PLoS Comput Biol 2006, 2, e27.
15. Dinner, A. R.; Lazaridis, T.; Karplus, M. Proc Natl Acad Sci USA 1999, 96, 9068.
16. Fuchs, P. F. J.; Bonvin, A. M. J. J.; Bochicchio, B.; Pepe, A.; Alix, A. J. P.; Tamburro, A. M. Biophys J 2006, 90, 2745.
17. Chen, J.; Im, W., Brooks C. L. J Am Chem Soc 2006, 128, 3728.
18. Chebaro, Y.; Dong, X.; Laghaei, R.; Derreumaux, P.; Mousseau, N. J Phys Chem B 2009, 113, 267.
19. Ishikawa, K.; Yue, K.; Dill, K. A. Protein Sci 1999, 8, 716.
20. Kaur, H.; Garg, A.; Raghava, G. P. S. Protein Pept Lett 2007, 14, 626.
21. Thomas, A.; Deshayes, S.; Decaffmeyer, M.; Van Eyck, M. H.; Charloteaux, B.; Brasseur, R. Proteins 2006, 65, 889.
22. Nicosia, G.; Stracquadanio, G. Biophys J 2008, 95, 4988.
23. Camproux, A. C.; Tuffery, P.; Chevrolat, J. P.; Boisvieux, J. F.; Hazout, S. Protein Eng 1999, 12, 1063.
24. Camproux, A. C.; Gautier, R.; Tufféry, P. J Mol Biol 2004, 339, 591.
25. Tuffery, P.; Derreumaux, P. Proteins 2005, 61, 732.
26. Tuffery, P.; Guyon, F.; Derreumaux, P. J Comput Chem 2005, 26, 506.
27. Maupetit, J.; Tuffery, P.; Derreumaux, P. Proteins 2007, 69, 394.
28. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
29. Frishman, D.; Argos, P. Proteins 1995, 23, 566.
30. Jagielska, A.; Scheraga, H. A. J Comput Chem 2007, 28, 1068.
31. Yang, J. S.; Chen, W. W.; Skolnick, J.; Shakhnovich, E. I. Structure 2007, 15, 53.
32. Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Nucleic Acids Res 1997, 25, 3389.
33. Suzek, B. E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C. H. Bioinformatics 2007, 23, 1282.
34. Derreumaux, P.; Mousseau, N. J Chem Phys 2007, 126, 025101.
35. Mousseau, N.; Derreumaux, P. Acc Chem Res 2005, 38, 885.

36. Song, W.; Wei, G.; Mousseau, N.; Derreumaux, P. J Phys Chem B 2008, 112, 4410.
37. Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L. Nucleic Acids Res 2008, 36(Database issue), D402.
38. Berjanskii, M. V.; Wishart, D. S. Nucleic Acids Res 2007, 35(Web Server issue), W531.
39. Jones, D. T. J Mol Biol 1999, 292, 195.
40. Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. Biophys J 2008, 94, L75.
41. Casarotto, M. G.; Gibson, F.; Pace, S. M.; Curtis, S. M.; Mulcair, M.; Dulhunty, A. F. J Biol Chem 2000, 275, 11631.
42. Siedlecka, M.; Goch, G.; Ejchart, A.; Sticht, H.; Bierzynski, A. Proc Natl Acad Sci USA 1999, 96, 903.
43. Rozek, A.; Friedrich, C. L.; Hancock, R. E. Biochemistry 2000, 39, 15765.
44. Tinoco, L. W.; Da Silva, A.; Leite, A.; Valente, A. P.; Almeida, F. C. L. J Biol Chem 2002, 277, 36351.
45. Pitera, J. W.; Swope, W. Proc Natl Acad Sci USA 2003, 100, 7587.
46. Kannan, S.; Zacharias, M. Proteins 2007, 66, 697.
47. Duan, Y.; Kollman, P. A. Science 1998, 282, 740.
48. Lei, H.; Deng, X.; Wang, Z.; Duan, Y. J Chem Phys 2008, 129, 155104.
49. Verma, A.; Gopal, S. M.; Oh, J. S.; Lee, K. H.; Wenzel, W. J Comput Chem 2007, 28, 2552.
50. Ding, F.; Tsao, D.; Nie, H.; Dokholyan, N. V. Structure 2008, 16, 1010.