JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Improved PEP-FOLD Approach for Peptide and Miniprotein Structure Prediction

Yimin Shen,[†,§,#] Julien Maupetit,[‡,§,#] Philippe Derreumaux,[‡,¶,§,∥] and Pierre Tufféry*[,†,§,#]

[†]INSERM U973, MTi, F-75205 Paris, France

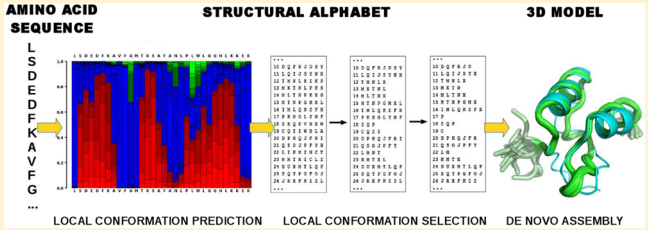
[‡]Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Institut de Biologie Physico-Chimique, F-75005 Paris, France

[¶]Institut Universitaire de France, 103 Boulevard Saint-Michel, 75005, Paris, France

[§]Univ Paris Diderot, Sorbonne Paris Cité, F-75205 Paris, France

Supporting Information

ABSTRACT: Peptides and mini proteins have many biological and biomedical implications, which motivates the development of accurate methods, suitable for large-scale experiments, to predict their experimental or native conformations solely from sequences. In this study, we report PEP-FOLD2, an improved coarse grained approach for peptide de novo structure prediction and compare it with PEP-FOLD1 and the state-of-the-art Rosetta program. Using a benchmark of 56 structurally diverse peptides with 25−52 amino acids and a total of 600 simulations for each system, PEP-FOLD2 generates higher quality models than PEP-FOLD1, and PEP-FOLD2 and Rosetta generate near-native or native models for 95% and 88% of the targets, respectively. In the situation where we do not have any experimental structures at hand, PEP-FOLD2 and Rosetta return a near-native or native conformation among the top five best scored models for 80% and 75% of the targets, respectively. While the PEP-FOLD2 prediction rate is better than the ROSETTA prediction rate by 5%, this improvement is non-negligible because PEP-FOLD2 explores a larger conformational space than ROSETTA and consists of a single coarse-grained phase. Our results indicate that if the coarse-grained PEP-FOLD2 method is approaching maturity, we are not at the end of the game of mini-protein structure prediction, but this opens new perspectives for large-scale in silico experiments.



INTRODUCTION

Fast and accurate peptide structure characterization remains a long-standing goal in structural biology and peptide engineering since peptides up to 50 amino acids represent a source of novel antibiotics and therapeutics.[1] In addition, these amino acid sizes can fold autonomously and be the functional centers of full length proteins (e.g., C1, UBA, and WW, to cite some).[2−4] One major obstacle in predicting peptide structures, in contrast to larger proteins, is that only a small number of solution structures have been characterized and are available in structural databases. On October 1st, 2013, the number of entries of the Protein Data Bank (PDB)[5] corresponding to isolated proteins of less than 51 amino acids was 2057, and only 799 proteins had less than 30% sequence identity and their structures not solved in a membrane environment. In addition, de novo sequences can deviate from those in the PDB by more than 70% sequence identity, making the use of comparative modeling techniques unreliable when no experimental information is available. For instance, it is remarkable that the de novo peptide with the helix-turn-helix motif designed in 2004 (PDB 1vrz) or with the beta-alpha-beta motif (PDB 2ki0) designed in 2009 still do not have any homologue in the PDB.

Considering the number of new sequences that are delivered by each genome project, we need to go beyond time-consuming simulations of all-atom systems in explicit solvent, though molecular dynamics studies show success in folding diverse structurally proteins with 10−80 amino acids by using the specially designed Anton computer[6] or the Folding-at-home project.[7] Present estimates of the number of hypothetical peptide coding sequences in the complete prokaryotic genomes available today are on the order of 1.5 million.[57] In eucaryotes, the number of peptide candidates is even higher, with estimates of the number of venom peptides on the order of 12 millions.[8] This highlights the need for fast approaches to model the structure of peptide and small proteins.

The most efficient and rapid methods are multiscale in character. Such methods start sampling with low resolution models, use fragment assembly (FA) methods and then select some conformations for subsequent full-atom refinements. These include the widely used Rosetta,[9,10] I-Tasser,[11] and Quark[12] methods. Other Web servers include PepStr,[13] Bhageerath,[14] and Peplook.[15] Other programs such as Zipping and Assembly,[16] the AWSEM-based approach,[17] the conformational space annealing,[18] GPS,[19] and replica exchange molecular dynamics simulations (REMD) with OPEP[20] are not open and

Received: April 5, 2014
Published: August 20, 2014

   dx.doi.org/10.1021/ct500592m | J. Chem. Theory Comput. 2014, 10, 4745−4758

Table 1. Peptide Collection[a]

no.	PDB	exp	BMRB	models	pH	topo	L	L3D	clust	core seq	no.	PDB	exp	BMRB	models	pH	topo	L	L3D	clust	core seq
1	1by0	NMR		20	6.8	a	27	27		1−23	29	1ywj	NMR	6558, 6559, 6719	15	6	b3	41	28		1−28
2	1yyb	NMR	6556	20	6.5	a	27	26		1−20	30	1ymz	NMR	6530	10	7	b3	43	37		3−13, 15−32
3	2kbl	NMR	16090	10	7	b2	29	27		6−27	31	2dmv	NMR		20	7	b3	43	43	2zaj	7−37
4	2k76	NMR	7232 7233, 15946	10	7	ba	30	30		4−29	32	2k9d	NMR	15993 15994, 16000	20	7	a3	44	44		1−10, 12−27, 29−44
5	2gdl	NMR		9	6	aca	31	31		8−8, 10−11, 14−15, 21−29	33	2p81	NMR	7386, 15520, 15536	25	5.7	a2	44	44		8−38
6	2l0g	NMR		20	7	a2	32	32		5−32	34	1f4i	NMR		21	6.5	a3	45	45	1dv0	2−40
7	2bn6	NMR		29	6.5	a2	33	33		4−29	35	1p9c	NMR		10	6.5	ca	45	45		16−33
8	2ovc	X-ray	n/a	1	5.5	a	33	30		1−30	36	1usd	X-ray	n/a	1	8.5	a	45	41	1use	1−41
9	1bwx	NMR	1666, 3427, 3449	10	5.8	a2	34	39		15−28	37	1vpu	NMR		9	7	a3	45	45		5−17, 23−29, 39−42
10	2kya	NMR	16182 16943	14	5.7	ac	34	34		11−30	38	3e21	X-ray	n/a	1	7.5	a3	45	40		1−40
11	1wy3	X-ray	n/a	1	7	a3	35	35		1−86	39	1pv0	NMR	5847	25	6.8	a2	46	46		3−44
12	1wr3	NMR		20	6.5	b3	36	36	2l4j	5−15,17−34	40	2e5t	NMR	11000	20	6.8	a2	46	46		2−19, 21−46
13	1wr4	NMR		20	6.5	b3	36	36	1ymz	5−34	41	2jnh	NMR	15111	15	6.5	a3	46	44		4−41
14	2ki0	NMR		20	6	bab	36	36		5−11, 13−36	42	2l4j	NMR		21	5.5	b3	46	46		12−39
15	1e0m	NMR	4713	10	6.5	b3	37	37	1ymz	3−35	43	1dv0	NMR	4757	18	6.5	a3	47	45		2−40
16	1e0n	NMR	4715	10	6.5	b3	37	27		1−25	44	1use	X-ray	5955	1	8.5	a	47	40		1−40
17	1yiu	NMR	6459 15153	8	6	b3	37	37	2l4j	5−36	45	1w4e	NMR		20	5.5	a2	47	45	1w4g	4−44
18	1bhi	NMR	4216	20	6.3	b2a	38	38		7−11, 14−33	46	1w4g	NMR		20	5.5	aca	47	45		4−44
19	1i6c	NMR	4882, 5248, 16070, 16088	10	6.4	b3	39	39	1ymz	2−11, 16−33	47	2btg	NMR		20	6.5	aca	47	45	1w4h	5−44
20	1jrj	NMR		36	5.9	a	39	39		11−38	48	2ekk	NMR		20	7	a3	47	47		8−46
21	2ysc	NMR		20	7	b3	39	39		8−39	49	2wxc	NMR		20	7	aca	47	47	1w4h	7−47
22	1e0l	NMR	4714,11007, 11008,15453	10	6.5	b3	40	37	1ywj	6−37	50	1ify	NMR		10	6.5	a3	49	49		5−48
23	2ysf	NMR		20	7	b3	40	40	2l4j	7−39	51	1nd9	NMR		10	6	ba2ba	49	49		1−42, 45−49
24	2ysg	NMR		20	7	b3	40	40	2ysh	7−38	52	2j8p	NMR		30	6	a3	49	49		2−44
25	2ysh	NMR		20	7	b3	40	40		7−37	53	2ysb	NMR		20	7	b3	49	49		9−43
26	2ysi	NMR		20	7	b3	40	40	1ywj	9−40	54	2zaj	NMR	10215	20	7	b3	49	49		9−44
27	1k1v	NMR		20	6.7	a3	41	41		1−41	55	1w4h	NMR	2546	20	7	aca	51	45		5−44
28	1wr7	NMR		20	5.7	b3	41	41	2l4j	9−40	56	1pgy	NMR	15968	20	6	a3	52	47		4−44

[a]For each, we report the Protein Data Bank identifier (PDB), the method used to resolve the structure, the BMRB identifier when availabole, the number of models (models), the pH at which the structure was determined, its topology (topo) in terms of alpha halix (a) or beta strand (b), the size of the sequence used for simulations (L) and that of the structure (L3D). Clust denotes the identifier of the representative of the homologues, when different from the peptide. Core seq denotes the residues of the rigid core, numbered from from 1.

dx.doi.org/10.1021/ct500592m | J. Chem. Theory Comput. 2014, 10, 4745−4758

accessible to the biologist community yet. In the FA method, the protocol is separated into two steps: first, one collects structural candidates for every short segment of the target sequence, retrieving them from the PDB; the second step is to assemble these fragments for constructing tertiary structures that have low energies. Using computer simulations, it has been shown that these local biases shape up the funnel free-energy landscape,[21] and this local structuring provides information to solve some but not all of the conformational search problems as demonstrated by REMD simulations of 872 8-mer, 12-mer, and 16-mer peptide fragments from 13 proteins using the AMBER96 force field and the OBC implicit solvent model.[22]

We have recently developed PEP-FOLD,[23,24] based on the concept of structural alphabet (SA) and the use of 27 letters to describe proteins as series of overlapping fragments of four amino acids.[25] PEP-FOLD uses a two-step procedure: (i) the prediction of a limited set of SA letters at each position from sequence using an amino acid profile as input of a Support Vector Machine (SVM) and (ii) the progressive assembly of the prototype fragments associated with each SA letter using a greedy algorithm and the sOPEP coarse-grained force field. Using a set of 25 peptides with 9−23 amino acids, the first version of PEP-FOLD (PEP-FOLD1)[23] identified, on average, lowest-energy conformations differing by 2.6 Å C_α RMS deviation (RMSD) from the Nuclear Magnetic Resonance (NMR) structures. For 13 mini-proteins with 27−49 amino acids, PEP-FOLD1 reached an accuracy of 3.6 and 4.6 Å RMSD for the most-native and lowest-energy conformations, using the nonflexible regions identified by NMR. Though limited to linear peptides of 10−25 amino acids, and recently updated to peptides of 10−36 amino acids with disulfide bonds,[26] PEP-FOLD1 has proven efficient for various applications. Several PEP-FOLD predictions have been supported by circular dichroism or small-angle X-ray scattering experiments (e.g., refs 27−29) or used for the design of immunogenic, antiviral, or antimicrobial peptides, the characterization of protein−peptide interactions, and the conformations of protein fragments (e.g., refs 30−37). However, for several mini-proteins, PEP-FOLD1 was not able to identify the native fold.

Here, we present PEP-FOLD2, which by revisiting the prediction of the SA letters from the sequence, allows a more accurate de novo structure prediction of miniproteins up to 52 amino acids. We compare its performances with PEP-FOLD1 and Rosetta predictions using a benchmark of 56 soluble miniproteins with 25−52 amino acids. Each system was subject to 600 PEP-FOLD1, PEP-FOLD2, and Rosetta simulations. Overall, with no experimental structures at hand, PEP-FOLD2 improves the predictions relative to PEP-FOLD1 and Rosetta by 12% and 5%. Yet, compared to Rosetta, this higher prediction rate is achieved by exploring a larger conformational ensemble, i.e. making the identification of near-native and native conformations more difficult.

MATERIALS AND METHODS

Data Sets. Our set of peptides/mini-proteins consists of 56 linear peptides in solution. It corresponds to an exhaustive survey of the protein data bank[5] in March of 2011, searching the pdb_seqres for sequences between 25 and 50 amino acids. Several filters have been applied to the collection of 331 peptides identified to discard peptides with disulfide bonds or non-natural amino acids, amyloid peptides, transmembrane peptides, peptides bound to membranes or ligands, and peptides structurally characterized at a pH less than 5.5 or in complexes or stabilized by metal ions. Note we kept the peptides differing at least by one amino acid to have information on the impact of limited sequence divergence on model stability. When possible, the NMR structure has been preferred. Finally, the sequences of the experimental structures have been compared to those of the pdb_seqres and of the BMRB,[38] and when different, the longest has been retained. The longest sequence in our set is thus of 52 amino acids. Details about the 56 peptides used as a benchmark are provided in Table 1. Sequence similarity has been assessed using kClust[39] specifying 60% coverage in the alignment, and using a cutoff of 2.33 which corresponds approximately to 50% sequence identity. The rigid core of the structures have been identified using the procedure described in our previous work.[24] It excludes residues considered as flexible, i.e. displaying over the NMR models, cRMSf values > 1.5 Å. For X-ray structures, only one model being provided, the rigid core corresponds to the complete structure. We emphasize that many structures were solved in aqueous solution with some buffers, containing 10−50 mM potassium phosphate (e.g., 2k76 and 2bkl) along with 100−200 mM NaCl (e.g., 1by0 and 1vpu) or even trifluoroethanol (TFE) (e.g., 2gdl and 2kya).

To learn the thresholds used to bias the conformational sampling and to avoid any bias due to learning from peptides or short proteins, we have used a nonredundant collection of proteins of more than 50 amino acids proposed by the culled PDB[40] on February 1st, 2010. This collection contained PDB entries solved at a resolution less than 2 Å, with an R value not more than 0.25 and was filtered at 40% sequence identity. Starting from 7087 chains, we have selected all the chains having no breaks and no hetero groups. Distinct learning and validation sets have been identified on the basis of the chain lengths, between 150 and 300 residues (S150−300, 1295 chains) and 50 to 150 residues (S50−150, 1304 chains), respectively.

Structural Alphabet Paradigm. PEP-FOLD relies on the concept of a structural alphabet (SA). SA can be assimilated to a generalization of the secondary structure, with a larger number of canonical conformations (SA states, or letters). Here, we used a Hidden Markov Model derived structural alphabet (HMM-SA).[25] In this model, a protein or peptide is considered as a series of fragments of four residues overlapping by three residues. Hence, a protein of L amino acids corresponds to a series of $L - 3$ fragments. The fragments are associated with four geometrical descriptors, namely the three distances between the nonconsecutive alpha carbons of the fragment, and the triple-product defining the signed volume of the fragment. The parameters of the HMM model are the mean values and covariance matrices of the descriptors of each state, and the transition matrix is associated with the first order Markovian process.[25] Given the model and the descriptors associated with a conformation of a protein of size L, it is possible to identify the series of the $L - 3$ SA letters that optimally describe the conformation using the Viterbi algorithm.[41] It is also possible to identify the probability that each letter emits each of the four amino acid fragment of the protein using the forward−backward algorithm.[42] In this study, we used a 27 letter SA found to correspond to a statistical optimum. Each SA state is associated with a limited number of representative fragments—or prototypes—that have been chosen to sample the conformational variability of the letter. In theory, the number of prototypes depends on the conformational variability of the letter and varies from 1 to

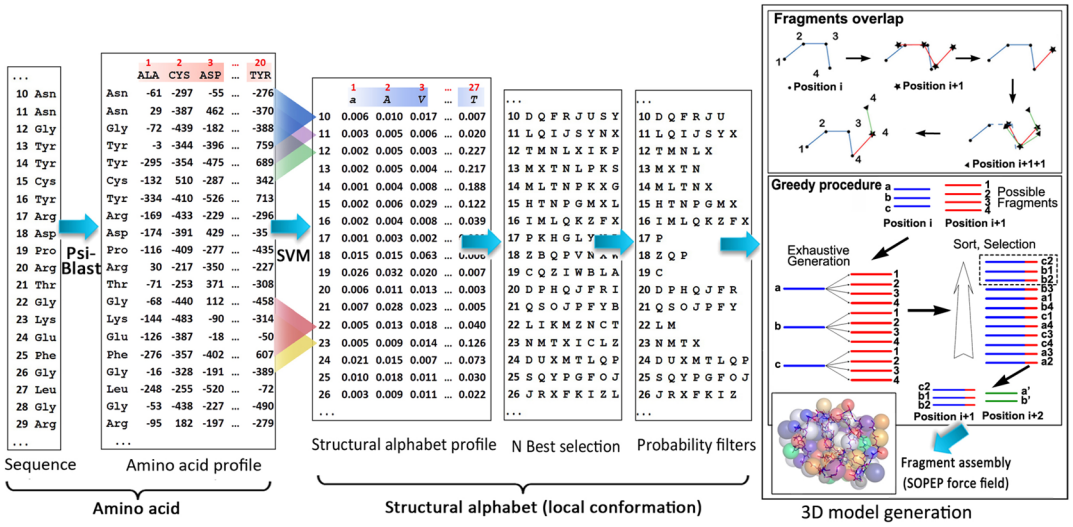


Figure 1. Flowchart of the 3D model generation procedure.

22. In practice, this number of prototypes is set to three for most SA letters and eight for two fuzzy letters.[24,43]

3D Model Generation Procedure. Figure 1 presents an overview of the 3D modeling procedure. It consists of three steps. The first step predicts from the amino acid sequence a profile of local conformations, called a SA profile. It describes the probability of each of the 27 SA letters for each consecutive $L - 3$ fragment in the sequence. In a second step, the SA profile is processed to select, for each position in the sequence, a limited number of SA letters, i.e. a limited number of local conformations. Finally, the fragments associated with the selected SA letters are assembled progressively to produce full structure models. We describe more in detail each step in the following sections.

Structural Alphabet Profile Generation. Given a sequence, we predict, for each of the $L - 3$ overlapping fragments in a sequence of size L, the probability that it has, in the protein structure, a conformation associated with each of the structural alphabet letter. The prediction approach used in this study is identical to that used by Maupetit et al.[24] Briefly, we have trained a support vector machine (SVM) to predict the probabilities of each state given a sequence of four amino acids enlarged by two amino acids each side[24]—i.e. using a window of eight amino acids centered on the four amino acids of interest. In practice, the SVM takes as input a matrix of 20×8 values, where each series of 20 values corresponds to the probabilities of the 20 amino acids at the corresponding positions, obtained using PSI blast[44] against the Uniprot collection filtered at 90% sequence identity.[45] We perform the prediction for each fragment k, in turn. The output of the prediction is thus, for a sequence of L amino acids, a SA profile of $L - 3$ vectors of 27 values, where vector k, $1 \leq k \leq L - 3$, corresponds to the 27 probabilities p_k^l that the kth fragment is associated with the letter l. Our SVM predictor has been trained using a collection of 3672 protein chains corresponding to a non redundant collection of proteins selected in 2006. In our experience, updates of this set have so far not led to significant improvement, which can be related to the locality of the prediction.

Posterior Analysis of the Structural Alphabet Profile. In the framework of our procedure, searching the complete conformational space is equivalent to accepting all the 27 SA letters for each fragment in the sequence. In a previous work,[24] we have shown that the first eight SA letters of highest probabilities for each fragment are sufficient to generate native models. In the present study, we have further investigated the probability values, by analyzing the cumulative and complementary cumulative distributions to identify probability thresholds below which we can discard an unlikely SA letter and above which we can consider a SA letter well predicted and retain only one letter, respectively for each fragment. We have performed this analysis for each SA letter and for the two groups of SA letters representative of α helix and extended conformations.

We consider as observations the encoding of the 3D protein structures as series of SA letters obtained using the Viterbi algorithm. We note $O_k = l$, where O stands for observation of the SA letter l obtained at position k in the encoding of a 3D structure. We analyze the SA profile generated from the sequence using a SVM. p_k^l is the estimated probability of l at position k, by the SVM. The distribution of estimated probabilities of letter l is denoted as $F(l)$. The distribution of the estimated probabilities of l over the positions where $O_k = l$ is denoted as $F^c(l)$, with $F^c(l) = \text{Prob}(p_k^l \leq x | O_k = l)$. The corresponding complementary distributions are $\overline{F} = 1 - F$ and $\overline{F}^c = 1 - F^c$.

For a given letter l, we have searched, over all positions k, the largest threshold value p_{min}^l, and the smallest threshold values p_{max}^l such as

$$\text{Prob}(O_k = l | p_k^l < p_{\text{min}}^l) \leq \epsilon \quad (1)$$

$$\text{Prob}(O_k \neq l | p_k^l \geq p_{\text{max}}^l) \leq \epsilon \quad (2)$$

where ϵ is the risk of error.

It is important to note that in some cases, considering the SA letter identity as stated by $O_k = l$ can be too stringent. Indeed, by analyzing collections of structural alignments, Guyon et al.[46] have observed that some equivalence—or confusion—between the structural alphabet letters can be accepted which can be formalized in terms of a similarity matrix. We denote l^+ the set of letters which may be confused with l, i.e. for which the substitution matrix has positive values. Similarly to p_{max}^l, it is possible to search for the smallest threshold value p_{max}^{l+} such that

$$\mathrm{Prob}(O_k \notin l^+ | p_k^l \geq p_{\max}^{l+}) \leq \epsilon \quad (3)$$

Finally, we have also considered the two groups of 5 SA letters occurring in the classical helical and extended secondary structures, denoted as SA_α and SA_β, respectively. For the exact details of the frequencies of occurrence of the SA letters in helical and extended secondary structures, see the work of Camproux et al.[25] The enlarged subsets of letters accepting confusion are denoted as SA_α^+ and SA_β^+. We have searched for the smallest $p_{\max}^{ss+}$ value such that

$$\mathrm{Prob}(O_k \notin \mathrm{SA}_{ss}^+ | \sum_{l \in ss} p_k^l \geq p_{\max}^{ss+}) \leq \epsilon \quad (4)$$

Correct Prediction Assessment. Having selected a limited number of SA letters at position k, which we note as $SA(k)$, a correct prediction corresponds to $O_k \in SA(k)$ and an error corresponds to $O_k \notin SA(k)$. CP is the ratio of good prediction. Considering confusion, SA_k^+ corresponds to SA_k extended by SA letters with positive values in the substitution matrix for at least one letter of SA_k, and a correct prediction is obtained when $O_k \in SA_k^+$. In the following, CP^+ denotes the ratio of good prediction accepting confusion. Using the secondary structure classes, a correct prediction is obtained when $(SA(k) \cap SA_{ss} \neq \varnothing)$ and $(O_k \in SA_{ss})$, where ss is one of α or β. Introducing the confusion, a correct prediction is obtained when $(SA_k^+ \cap SA_{ss} \neq \varnothing)$ and $(O_k \in SA_{ss})$. The corresponding ratios of correct prediction are denoted by CP_{ss} and CP_{ss}^+, respectively.

3D Model Generation Using a Polypeptide Chain Growth Approach. Having identified $SA(k)$, we use a greedy procedure to grow the polypeptide chain by adding one residue at a time. Each SA letter is associated with a limited number of representative fragments—or prototypes—that have been chosen to sample the conformational variability of the letter (see the section about the structural alphabet paradigm). The elementary assembly process is achieved by overlapping the three first residues of a prototype on the last three residues of the polypeptide chain grown so far. Since our procedure relies on a limited number of prototypes for each SA letter, the assembly is not performed in a continuous, but rather in a discrete space. To sample the combinatorial of the assembly, all the representative fragments of the SA letters in SA_k are used to generate conformations increased by one residue. The energies of the conformations are calculated using the sOPEP force field.[47] Since the number of conformations to generate the complete structure is too large to be systematically explored, PEP-FOLD uses a stochastic selection procedure to prune it (see refs 24 and 43), retaining 3000 conformations of both lower energy and randomly chosen conformations for the next iteration. More details about the model generation can be found in ref 43. The growing process can start at any position of the structure. The procedure then alternatively adds residues at each side of the growing structure. We found more efficient not to start from regions presumably in secondary structures and, for peptides with more than 30 amino-acids, not to start from the eight residues of the N terminus and C terminus. Once the full structure is constructed, the model is refined by a Monte Carlo procedure of 300 000 steps, using a random selection of one prototype at each step, so as to optimize the conformation of the complete structure. Finally, since the procedure contains some stochastic aspects, several runs are performed. For this study, we used three series of 200 independent runs, each series based on a different conformational bias (see Results and Discussion).

Model Comparison with Experimental Conformations. While one important aspect of CASP and numerous studies was to identify the right test measures to compare models and experimental conformations—including the RMSD, the GDT_TS,[48] and the TMscore[49] among others—most of them are satisfactory only for large systems. The well-known dependence of the RMSD on protein size makes uneasy to identify cutoff RMSD values representative of a significant similarity. The widely used TMscore[49] has has been calibrated for proteins with more than 80 amino acids, but it can be non relevant for shorter sizes. In this work, we used the recently developed BCscore[50] found more discriminating than the TMscore to identify non native models for peptides and miniproteins. The BCscore varies between 1 (perfect match) and −1 (perfect mirror). Values around 0 correspond to unrelated conformations. Models with BCscore $\leq$ 0.6 and BCscore $\geq$ 0.8 can be considered as incorrect or native, respectively. A model with 0.6 $\leq$ BCscore $\leq$ 0.8 can be considered as near-native, i.e. only partly consistent in terms of secondary structures and topologies.

Model Clustering and Ranking. To cluster the models, we have used a standard clustering procedure using as distance $d = 1 - \mathrm{BCscore}$, and a cutoff value of 0.2, i.e. the value corresponding to a BCscore of 0.8. Once the clusters defined, they can be ranked according to their energy, based on the lowest sOPEP energy over the models of the cluster. We have also used Apollo[51] as an alternative approach to score the quality of each model. Apollo is based on the prediction of structural features from the sequence and it quantifies the agreement of the models with these predictions in terms of a predicted TMscore. Clusters can then be ranked according to their Apollo scores (each cluster is associated with the best predicted score over the models of the cluster).

To quantify the diversity of the conformations generated over series of simulations, we have used the so-called equivalent number of conformations (nEq), a measure of the conformational diversity based on the number of clusters and the fraction of models belonging to the cluster relative to the total number of models generated:

$$\mathrm{nEq} = \exp(H) \quad \text{with } H = -\sum_{i=1}^{n_C} f_i \log(f_i) \quad (5)$$

where n_C is the number of clusters identified, f_i corresponds to the fraction of models generated belonging to the cluster i, and H corresponds to the Shannon entropy of the distribution of the clusters. nEq varies from 1 to n_C depending on the cluster frequencies.

Rosetta Simulations. Rosetta simulations were performed using the standard implementation, starting from the fragments returned by the Robetta server.[9] As for each PEP-FOLD protocol, we performed three series of 200 Rosetta simulations for each target.

RESULTS AND DISCUSSION

Biasing the Conformational Space Sampling. In our previous studies, we have shown that it is possible to limit the conformational space to that described by the eight SA letters of highest probabilities at each position. Here, we have analyzed in a more accurate manner the predicted probabilities.

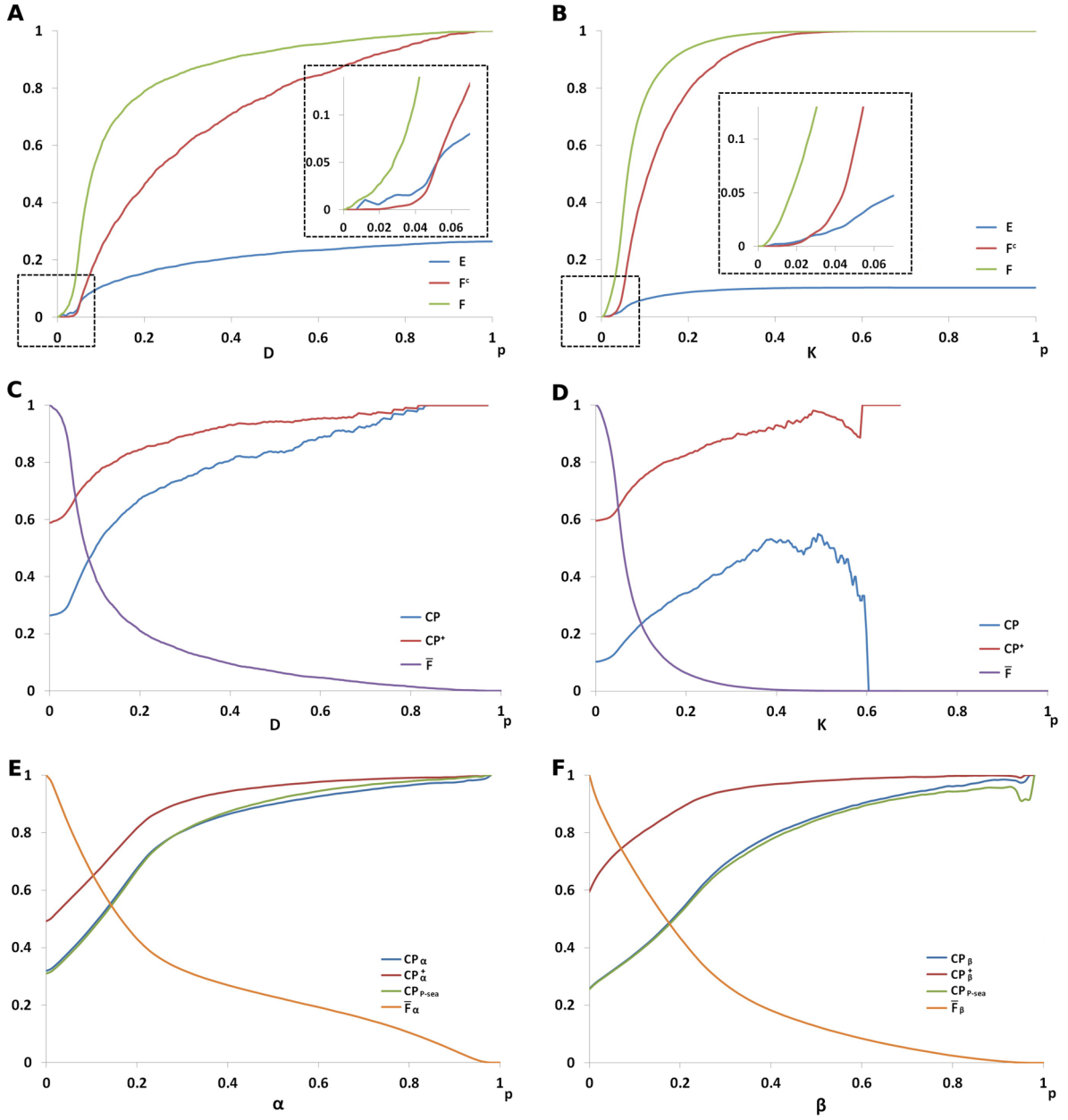


Figure 2. Predicted probability distributions to derive bias thresholds. Panels A and B depict for SA letters D and K (x axis labels) the distributions (F), conditional distributions (F^c) of the predicted probabilities, and the error rate (E). The complementary distributions ($\overline{F}$) and the rate of correct prediction (CP) and correct predictions accepting confusion (CP^+) are depicted panels C and D. Panels E and F depict complementary distributions of the classes of SA letters associated with α and β secondary structure ($\overline{F}_{ss}$) and the rate of correct prediction (CP_{ss}) and correct predictions accepting confusion (CP^+_{ss}). $CP_{p\text{-}sea}$ corresponds to the correct prediction taking the secondary structures as assigned by p-sea from the experimental structures instead of the 5 SA letter classes.

Figures 2A and B depict, for letters D (coil) and K (extended), the cumulative distributions (F), the conditional cumulative distributions (F^c), and the error rates - removal of a SA letter of probability lower than a threshold when is corresponds to the real letter in the structure, as a function of the probabilities predicted by the SVM. The same trends are observed for all SA letters (not shown). The cumulative distributions tend to increase rapidly, being over 0.5 for probability values as small as 0.1. Compared to the cumulative distributions, the conditional distributions tend to increase less rapidly, as expected since the values of the predicted probabilities tend to be larger for positions where the SA letter corresponds to the observation. The error rates are much smaller. Their maximum values correspond to the frequency of occurrence of the SA letters. Due to the weak frequencies of the SA letters, their variation also appears noisy for small probabilities, which makes difficult their use to directly identify a probability threshold below which one SA letter can be discarded at a given risk of error. To overcome this difficulty, we have considered the possibility to estimate the p^l_{min} values from the conditional cumulative distributions, which in our observations (see the insets of Figure 2A and B) are both a reasonable approximation and overall, an overestimation of the error for low probabilities. On average, we find that p^l_{min} values identified for a risk of error of 5% using the conditional cumulative distributions correspond to an error of less than 2%.

dx.doi.org/10.1021/ct500592m | J. Chem. Theory Comput. 2014, 10, 4745−4758

Figure 2C and D shows, for the same letters D and K, the complementary cumulative distributions ($\overline{F}$) and the correct prediction rates—not accepting (CP) or accepting some equivalence between the letters (CP^+). Similarly to the observations made for low probability values, the difference of the shapes of the CP curves are related to the low number of observations associated with large probability values. For instance, for SA letter K, there are only five p_k^K values larger than 0.6, among which only one corresponding O_k is K—correct prediction. And 191 p_k^K have a value of more than 0.4, among which 99 only correspond to correct predictions. However, this poor performance is only apparent. Accepting confusion between the SA letters, there are in fact 178 over these 191 positions for which $O_k \in K^+$. Similarly, for all five positions where $p_k^K \geq 0.6$, we have $O_k \in K^+$. Following, the shapes observed for CP^+ appear more compatible with the identification of thresholds above which the SA letter can be considered as corresponding to the true observations given a risk of error. We have thus used the CP^+ data to estimate, for each SA letter, p_{max}^{l+}, the threshold above which the prediction can be considered as correct at a given error risk.

Considering the two groups of five SA letters representative of the alpha helices and the beta strands, Figure 2E and F depicts the complementary cumulative distributions ($\overline{F}_{ss}$) and the correct prediction, accepting or not equivalences between letters, for the α helix and beta strand classes (CP_{ss} and CP_{ss}^+). Since SA letters describe the conformation of fragments of only four amino acids, it is possible to observe distorted α-helical conformations in small loops and extended conformations in larger loops. In particular, some helical conformations can be similar to type II beta turns. To limit the impact of such events, we have smoothed the data using a sliding window. Here, we have transformed p_k^l into $w1p_{k-1}^l + w2p_k^l + w3p_{k+1}^l$, where $w1$, $w2$, and $w3$ are weights applied to the $k-1$, k and $k+1$ positions, respectively. We have found that using (3, 5, 3) as values for ($w1$, $w2$, $w3$), the CP_{ss} evolution is very similar to that obtained using p-sea,[52] a consensus approach to assign secondary structures from the 3D structures, instead of the structural alphabet letter classes. Consistently with our processing at the level of individual SA letters, we have chosen to use the CP^+ curves to derive p_{max}^{ss+} thresholds above which the prediction secondary structure class can be considered as correct at a given error risk. We have not considered the converse p_{min}^{ss} thresholds since they can lead to the erroneous elimination of some local conformations in loops.

Finally, we have created a procedure combining the different thresholds to bias the conformational search. Starting from the initial subset of eight best SA letters accepted to describe the possible conformations at position k, the p_{min}^l thresholds have been used to discard SA letters with low probabilities, the p_{max}^{l+} thresholds have been used to discard all SA letters but the one associated with a probability larger than the threshold, and the p_{max}^{ss+} have been used in a similar manner to discard all SA letters not belonging to the class of SA letters representative of the secondary structures. Using the threshold values associated with a risk of 2% and 5% for p_{max}^{l+} and p_{min}^l, respectively, we have observed a diminution of the number of SA letters to 6.3 and 6.1 on average. Combining the two filters, the resulting number of letters to only 5.1, i.e. a bias increase of 36%. The same analysis using p_{max}^{ss+} values for a risk of 2.5% has resulted in a reduction to 7.4 letters for each of the upper bounds, for alpha and beta classes, respectively. Finally, combining all the filters together, we have found our procedure is able to limit the number of SA letters to only 4.9 and 4.8, as average on the learning and validation sets, respectively, i. e. a reduction of over 40% compared to the eight initial values. Figure 3 shows, for each SA letter, the conformational bias brought by the p_{min}^l, p_{max}^{l+}, and p_{max}^{ss+} filters, respectively, on the validation set.

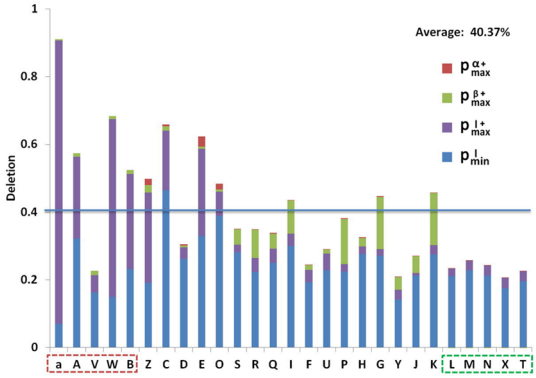


Figure 3. Conformational bias introduced by different filters. For the 27 SA letters, the effect of the consecutive filters are cumulated (bottom-up), and expressed in the fraction of SA letters discarded compared to the starting subset of eight letters at each position. The SA letters are sorted from most helical (left) to most extended (right). The five left letters surrounded in red correspond to the α helix class, and those in green (right) to the extended (beta) class. The terms p_{min}^l, p_{max}^{l+}, $p_{max}^{\alpha+}$, and $p_{max}^{\beta+}$ correspond to applying the cutoff values defined by eqs 1, 3, and 4, respectively.

It is important to assess if the reduced conformational space after filtering is compatible with the native conformation. Given the subset of SA letters selected at each position in a structure, this can be done, by checking if the correct SA letter or an equivalent SA letter if we accept confusion is consistent with the experimental conformation. Using the validation set, Figure 4 presents the correct prediction results detailed for each SA letter. PEP-FOLD1 results are obtained selecting at each position the eight SA letters with the largest probabilities. PEP-FOLD2 results are obtained considering the reduced subset obtained after filtering with cutoff values of 2, 5, 2.5, and 2.5% for p_{max}^{l+}, p_{min}^l, $p_{max}^{\alpha+}$, and $p_{max}^{\beta+}$, respectively. Ignoring any equivalence between the SA letters, we observe a decrease of the correct prediction from 83.8% down to only 65.8%. This decrease is compensated, however, by the selection of equivalent conformations, the difference between the correct prediction accepting confusion being on the order of only 2%. Overall the correct prediction score remains very high, close to 97.3% for the learning and validation sets in PEP-FOLD2 vs 99.5% in PEP-FOLD1. This is a good trade-off between errors and reduced numbers of letters considered.

Looking in detail at the score of each SA letter, the error rate appears rather evenly distributed, at the exception of the SA letter with label "O", which is the less frequent letter. Also, the SA letter with the label "a", observed in distorted helices, is the one for which the filters favor most of the time equivalent letters. It is important to note that an increased error rate of 2% implies on average one error for a sequence of 50 amino acids, but this still does not prevent the generation of a conformation close to the experimental one. Using our set of 56 peptides, we find that the filtered set of SA letters is able to approximate the experimental conformation at 1.7 Å RMSD, which remains close enough to the value of 1.1 Å RMSD obtained from the

dx.doi.org/10.1021/ct500592m | J. Chem. Theory Comput. 2014, 10, 4745−4758

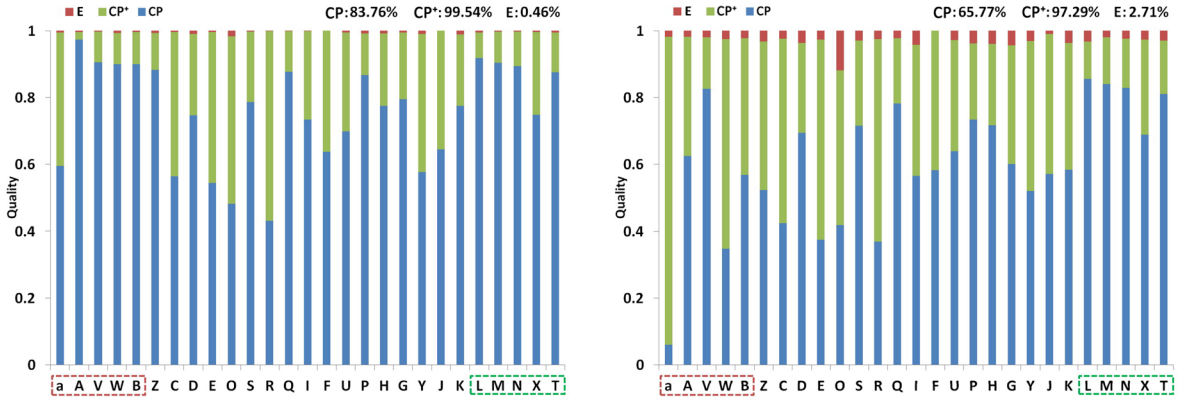


Figure 4. Correct prediction. The correct prediction rates are detailled for each SA letter. CP: the subset of local conformations selected contains the experimental one. CP*: the subset of local conformations selected contains a SA letter equivalent to the experimental one accepting confusion. *E*: error, i.e. remaining cases. (left) PEP-FOLD1. Eight SA letters are selected per position. (right) PEP-FOLD2. Here, 4.8 SA letters are selected per position on average.

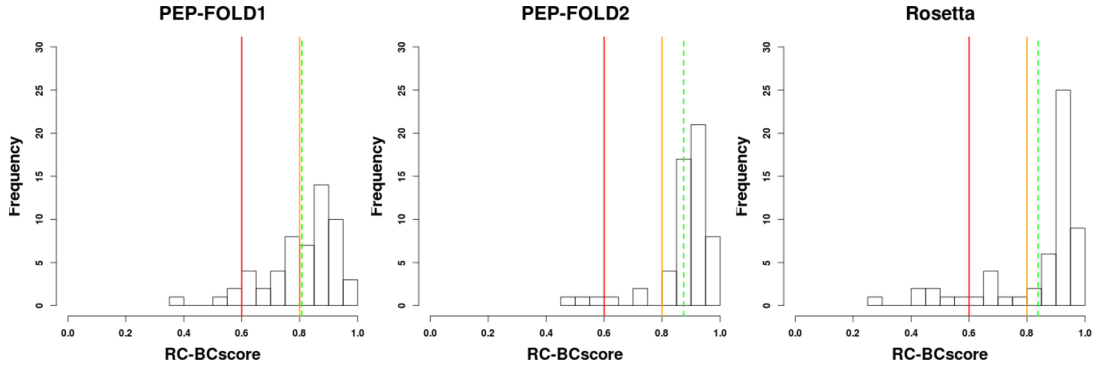


Figure 5. Best models. Deviation of the best models generated by PEP-FOLD1, PEP-FOLD2, and Rosetta from the experimental conformations. The deviation is expressed using the BCscore (1 stands for perfect match, 0 for fully uncorrelated conformations, −1 for mirror), calculated over the ridig cores of the structures (RC-BCscore). Vertical lines for BCscore values of 0.6 (in red) and 0.8 (in yellow) delimit the non native conformations (BC < 0.6), the near-native conformations (0.6 < BC < 0.8), and native conformations (BC > 0.8). The dashed line corresponds to the average BCscore calculated by using all 56 targets.

unfiltered set, and below the 2 Å conformational flexibility observed in molecular dynamics simulations.

Overall, while the 27 SA letters describe the full conformational space, and the PEP-FOLD1 strategy considers the eight best ranked SA letters per site, i.e. only 30% of the full space, our present results indicate that it is possible to bias even more the search. In this study we have set up, for the 3D reconstruction using PEP-FOLD2, a protocol of 200 runs combining three bias levels. 80 runs are based on the maximally biased trajectories (cutoff values of 2, 5, 2.5, and 2.5% for p^{l+}_{max}, p^{l}_{min}, $p^{\alpha+}_{max}$, and $p^{\beta+}_{max}$, respectively, i.e. an average of 4.8 SA letters per site reducing the size to be explored to 18%), 80 are based on moderately biased trajectories (cutoff values of 1, 0.1% for p^{l}_{min} and p^{l+}_{max}, respectively, no filter based on $p^{\alpha+}_{max}$ and $p^{\beta+}_{max}$, i.e. an average of 6.7 SA letters per site reducing the size to be explored to 25%), and 40 are based PEP-FOLD1 trajectories—minimal bias (eight SA letters per site).

De Novo Modeling of Soluble Peptides. Figure 5 shows, for all the 56 peptides of our test, the quality of the best models (largest BCscores with respect to the experimental structures) generated by PEP-FOLD1 and PEP-FOLD2. Since peptides can be flexible, the comparison is only made by over the rigid cores (see Materials and Methods).

A first observation is that PEP-FOLD2 generates higher quality models than PEP-FOLD1. The average BCscore of the best generated models increases from 0.80 (PEP-FOLD1) to 0.87 (PEP-FOLD2). This improved prediction rate does not come from a reduction in the number of incorrect models. By using a BCscore less than 0.6 for incorrect models, PEP-FOLD1 fails for four targets: 1vpu, 2gdl, 2ovc, and 2kya with BCscores of 0.55, 0.38, 0.52, and 0.57, respectively, while PEP-FOLD2 fails for only 3 targets: 2gdl, 2ovc, and 2kya (BCscores: 0.45, 0.52, and 0.56, respectively). The improvement for 1vpu is however limited (BCscore of 0.63, see below). The models of three problematic targets using PEP-FOLD1 are depicted in Supporting Information Figure 1. We see that the 31-residue 2gdl peptide has a rather unstructured NMR conformation with two α-helical regions that PEP-FOLD2 only partially generates. The 33-residue 2ovc peptide has a single helix X-ray conformation with a non helical C-terminus extremity that is modeled as helical by PEP-FOLD2. For the 34-residue 2kya peptide, the rigid NMR core consists of one helix for which PEP-FOLD2 unrolls eight residues at the C terminus.

A second observation is that PEP-FOLD2 leads to a major improvement for the targets that were determined as near-native by PEP-FOLD1, i.e. with a BCscore between 0.6 and 0.8. The number of such targets decreases from 22 with PEP-FOLD1 to only three with PEP-FOLD2: 2kbl, 1vpu, and 1nd9. The 29-residue 2kbl peptide displays a beta-hairpin by NMR; the PEP-FOLD2 models have the correct topology with a

different registry of hydrogen bonds. For the 45-residue 1vpu peptide, only two of the three helices of the rigid core are well folded, and their relative orientations differ from that in the NMR conformation. The 49-residue 1nd9 peptide is the target with the most complex topology of our test set. All the secondary structures are well identified, but the packing is not perfect.

Figure 6 shows two targets for which PEP-FOLD2 brought a significant improvement. The first is a 46-residue helical system

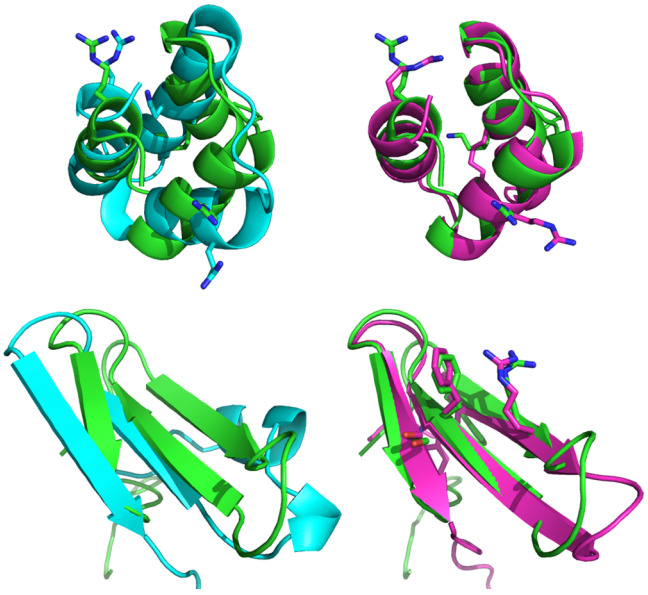


Figure 6. Best models using PEP-FOLD1 and PEP-FOLD2. Models generated for 2jnh (top) and 1i6c (bottom). The models generated by PEP-FOLD1 and PEP-FOLD2 in cyan and magenta, respectively. The experimental conformation is in green. Left and right images have identical orientations. The BCscores are of 0.78 and 0.97 for 2jnh and of 0.71 and 0.85 for 1i6c.

(2jnh) for which the NMR helix packing is improved. The second is a 39-residue three-stranded beta-sheet (1i6c) for which PEP-FOLD1 erroneously predicts a helix—though with a correct orientation relative to the sheet—instead of a third beta-strand.

It is interesting to assess how the three levels of bias are compatible with the generation of the best models. Using three series of 200 simulations on all targets, we find that the best models result from 50, 29, and 21% of the runs with the maximally, moderately, and minimally biased trajectories, respectively. We also find that excluding the models resulting from the PEP-FOLD1 minimally biased trajectories affects only marginally the quality of the overall results, with a decrease of BCscore of 0.02 on average. Over the series of runs, the results are systematically degraded for only four targets. These targets are 1yyb, 2kya, 2p81, and 2e5t, for which the decrease in BCscore are, on average, of 0.18, 0.19, 0.15, and 0.08, respectively. This does not prevent the generation of the native conformation for 2p81 and 2e5t, and 2kya is one target for which no native model could be generated even using the minimally biased trajectories. We thus find there is only one target (1yyb) for which only a near native conformation could be generated using the more biased trajectories. This target has an incorrect conformation for the first helix turn (see Supporting Information Figure 2).

Overall, the main finding of our analysis is that PEP-FOLD2 by introducing more local biases makes the energy landscape of miniproteins more funnel-like and facilitates the generation of native-like conformations, consistent with the results of ref 21. In addition, this more funnel-like characterization of the energy landscape compensates the small increased error rate of 2% in the selection of the letters and, in turn, increases the number of models of higher quality, with BCscores more than 0.8.

Finally, we turn to the ability of PEP-FOLD2 to recognize the most native conformations among all generated models with no experimental structures at hand. Figure 7 shows that the criterion of lowest-energy is not sufficient, with a clear decrease of the average BCscore of the models of lowest energy (average of 0.45). We recall that PEP-FOLD2 relies on a coarse grained energy force field and does not include the contribution of conformational entropy. Interestingly, models matching the experimental conformations can be identified in the best ranked suboptimal solutions. By considering the centroids of the first five clusters of lowest energy, the average BCscore is 0.68. A similar performance is obtained using Apollo (0.69). Overall, by using these five models, PEP-FOLD2 is able to propose near-experimental or experimental-like models for 45 targets among 56 (BCscores >0.6) and experimental-like models for 25 targets among 56 (BCscores > 0.8), when PEP-FOLD1 was ony able to propose near-experimental or experimental-like models for 38 targets and experimental-like models for 13 targets.

In addition to the targets 2gdl, 2ovc, and 2kya that are problematic for PEP-FOLD2, the other eight targets that are not ranked in the first five clusters are 1wg4 (ranked sixth), 1pvo (ranked eighth), 1jrj, 1use, and 1ify (ranked between 10th and 20th), and 2kbl, 1bhi, and 1vpu (ranked above 40th). For these last three targets, the variation of the Apollo scores among the clusters is very small −0.006, 0.03, and 0.028, respectively, which indicates a poor discrimination capacity. It is of interest to note that 1use and 1w4g are homologues to 1usd and 1w4e, respectively, i.e. targets for which near native or native conformations are ranked in the five best clusters. For 1w4g, the difference with 1w4e is the single substitution Y138W, and the 1w4g model, ranked sixth, has a near native topology (BC score 0.73). For 1use, the sequence difference with 1usd comes from the N terminus extremity (2 residues added), the BC score variation is associated with this region. For other pairs of close homologues such as 1wr3−1wr4, no impact is observed in terms of the identification of the native models.

Comparing with Rosetta. To assess PEP-FOLD2 efficiency, we now compare our results with those obtained by the state-of-the-art Rosetta program. We recall that we use 600 Rosetta simulations for each peptide. The results obtained for each target using PEP-FOLD2 and Rosetta are provided in Table 2, and Figure 8 depicts the models obtained by both methods for some targets.

Figure 5 shows the distribution of the best Rosetta models (largest BCscores with respect to the experimental structures) using our benchmark of 56 peptides. The average BCscore with Rosetta is 0.83, slightly less than the PEP-FOLD2 value of 0.86. Interestingly, PEP-FOLD2 and Rosetta do not perform well on the same peptides. First, the number of targets for which Rosetta fails to generate any near-native and native models is 7 (vs 3 with PEP-FOLD2). These 7 targets includes the three targets missed by PEP-FOLD (2kbl, 1vpu, and 1nd9) and the 1by0, 1usd, 1use, and 1yyb targets. The models obtained for these targets are depicted in Supporting Information Figure 1.

dx.doi.org/10.1021/ct500592m | J. Chem. Theory Comput. 2014, 10, 4745−4758

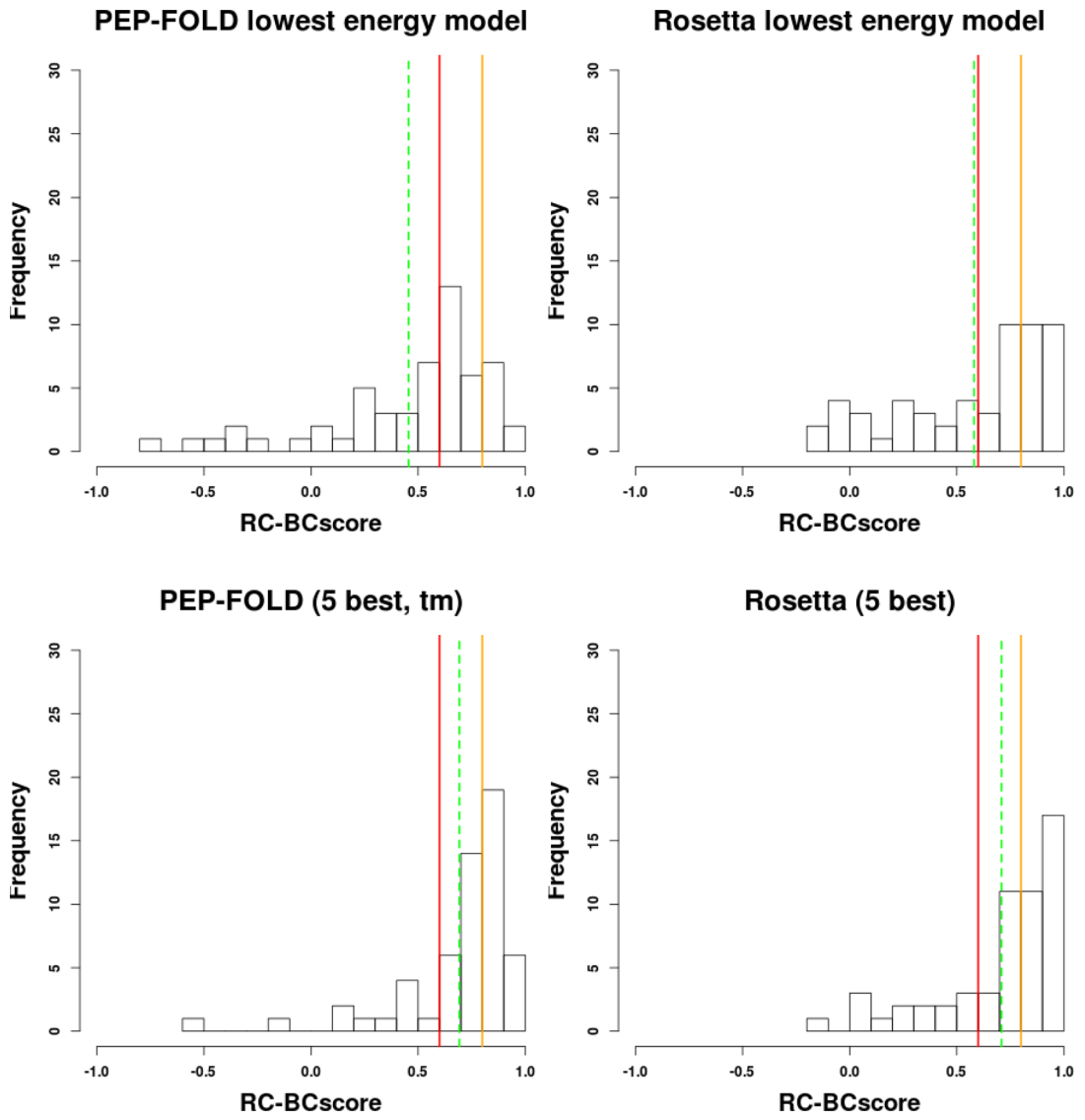


Figure 7. Best models returned by PEP-FOLD2 and Rosetta. (top) Lowest energy models (sOPEP and Rosetta score, respectively). (bottom) Best of five models identified using PEP-FOLD ranking procedure and Rosetta score, respectively. For the colors of the vertical lines, see Figure 5.

The 27-residue 1yyb by NMR consists of one α helix followed by a less structured Cter region for which Rosetta incorrectly fold the Nter residues as non helical. The 27-residue 1by0 by NMR, and the 45-residue 1usd and 47-residue 1use peptides by X-ray display one long α helix, while Rosetta incorrectly unfolds the helix C terminus for 1by0 and incorrectly breaks the helix for the two homologues 1usd and 1use. Note however that the 1usd and 1use structures have been solved by X-ray diffraction and their native structure consists of a coiled-coil tetramer domain stabilized by interpeptide interactions. To which extent interpeptide contacts stabilize the helical conformation of the monomer remains to be determined, but we found that the 2D predictions methods Porter[53] and Psi-pred[54] predict a long unbroken α helix, consistent with PEP-FOLD2 prediction.

Second, one notes that the number of very high quality models (BCscore > 0.9) obtained using Rosetta is slightly larger than that obtained with PEP-FOLD2-34 versus 29. This can be related to the fact that PEP-FOLD2 does not include a procedure similar to Rosetta's with multiple coarse-grained procedures and a final all-atom refinement procedure.

Looking at the ability of Rosetta to propose the experimental models with no experimental data at hand, Figure 7 shows that the all-atom Rosetta score performs only moderately better than the coarse-grained PEP-FOLD force field. The average BCscore of the lowest Rosetta score models is 0.58 vs 0.45 with PEP-FOLD2 but well below the average BCscore score of 0.86 for the best models. Similarly to PEP-FOLD2, native-like and native models can be identified in the best five ranked suboptimal solutions, the average BCscore reaching 0.70 with Rosetta vs 0.68 with PEP-FOLD2.

Overall from the selection of the five models of lowest energy, Rosetta is able to identify near-native or native models for 42 targets and native models for 28 targets. The Rosetta score fails to identify near-native models for 1by0, 1yyb, 2kbl, 2k76, 2gdl, 2ovc, 2kya, 1jrj, 2k9d, 1usd, 1vpu, 1use, 1w4e, and 1w4g, seven of which—2kbl, 2gdl, 2ovc, 2kya, 1jrj, 1use, 1w4g—are also incorrectly identified by PEP-FOLD2. The other targets include the pairs of homologues 1use−1usd and 1w4e−1w4g, 1yyb, and 1by0 for which Rosetta failed to generate near native models, 2k76 a miniature protein including

dx.doi.org/10.1021/ct500592m | J. Chem. Theory Comput. 2014, 10, 4745−4758

Table 2. Results per Target Using PEP-FOLD and Rosetta[a]

	PEP-FOLD				Rosetta			
PDB	best	lowE	best five	nEq	best	lowE	best five	nEq
1by0	0.86	0.65	0.68	4.4	0.40	0.65	0.05	39.6
1yyb	0.87	0.73	0.76	2.5	0.52	0.73	−0.12	29.2
2kbl	0.71	0.37	0.44	124.9	0.69	0.37	0.43	3.6
2k76	0.88	0.26	0.81	1.3	0.69	0.26	0.55	8.1
2gdl	0.45	0.05	0.15	13.1	0.41	0.05	0.29	57.2
2l0g	0.95	0.42	0.95	2.9	0.98	0.42	0.94	1.9
2bn6	0.86	0.58	0.80	4.4	0.91	0.58	0.87	5.4
2ovc	0.52	0.49	0.45	11.3	0.25	0.49	0.16	43.9
1bwx	0.93	0.70	0.78	11.2	0.90	0.70	0.63	8.5
2kya	0.56	0.26	0.26	92.2	0.45	0.26	0.21	21.9
1wy3	0.91	0.83	0.61	1.9	0.97	0.83	0.96	2.4
1wr3	0.94	0.80	0.88	24.7	0.92	0.80	0.89	1.2
1wr4	0.94	0.91	0.89	27.4	0.93	0.91	0.90	1.2
2ki0	0.89	0.63	0.80	5.3	0.94	0.63	0.90	1.0
1e0m	0.88	0.66	0.78	35.5	0.90	0.66	0.87	1.1
1e0n	0.92	0.68	0.84	85.8	0.95	0.68	0.92	1.0
1yiu	0.84	0.64	0.73	21.8	0.91	0.64	0.83	1.1
1bhi	0.82	−0.28	0.32	165.2	0.93	−0.28	0.88	162.3
1i6c	0.85	0.69	0.78	43.7	0.83	0.69	0.68	1.3
1jrj	0.85	0.11	0.55	4.3	0.63	0.11	0.42	10.2
2ysc	0.86	0.64	0.77	166.5	0.85	0.64	0.79	1.4
1e0l	0.92	0.28	0.85	70.0	0.89	0.28	0.78	5.8
2ysf	0.93	0.60	0.86	83.5	0.92	0.60	0.88	1.1
2ysg	0.93	0.35	0.90	48.0	0.91	0.35	0.83	2.8
2ysh	0.89	0.20	0.78	33.5	0.92	0.20	0.87	1.2
2ysi	0.85	0.78	0.77	50.1	0.79	0.78	0.75	1.3
1k1v	0.92	0.02	0.89	3.7	0.94	0.02	0.90	2.5
1wr7	0.95	0.88	0.94	68.0	0.93	0.88	0.91	1.2
1ywj	0.91	0.86	0.88	119.4	0.88	0.86	0.83	3.2
1ymz	0.93	0.81	0.88	49.0	0.94	0.81	0.92	3.0
2dmv	0.93	0.75	0.86	114.4	0.93	0.75	0.91	6.0
2k9d	0.82	0.60	0.72	5.5	0.68	0.60	0.57	2.2
2p81	0.96	0.59	0.90	14.0	0.93	0.59	0.74	4.3
1f4i	0.95	−0.49	0.90	5.6	0.98	−0.49	0.98	1.1
1p9c	0.86	0.52	0.64	37.1	0.86	0.52	0.74	137.9
1usd	0.87	0.30	0.79	3.6	0.47	0.30	0.35	15.6
1vpu	0.63	−0.57	−0.56	47.6	0.70	−0.57	0.55	3.5
3e21	0.91	0.78	0.76	4.4	0.99	0.78	0.98	2.1
1pv0	0.95	−0.39	0.46	6.2	0.94	−0.39	0.94	4.5
2e5t	0.92	0.42	0.71	2.9	0.97	0.42	0.97	7.0
2jnh	0.97	0.75	0.75	15.2	0.94	0.75	0.72	1.9
2l4j	0.90	0.69	0.85	108.7	0.92	0.69	0.90	2.3
1dv0	0.94	0.83	0.75	3.5	0.98	0.83	0.98	1.4
1use	0.87	0.50	0.49	7.8	0.55	0.50	0.31	45.6
1w4e	0.94	0.69	0.65	91.3	0.94	0.69	0.05	4.6
1w4g	0.90	−0.06	0.10	48.9	0.91	−0.06	0.01	2.9
2btg	0.96	0.91	0.83	36.4	0.94	0.91	0.79	1.8
2ekk	0.97	0.51	0.86	7.2	0.98	0.51	0.98	5.5
2wxc	0.91	0.51	0.83	36.5	0.90	0.51	0.73	2.5
1ify	0.85	−0.35	−0.16	8.7	0.97	−0.35	0.95	1.6
1nd9	0.73	0.64	0.65	5.6	0.69	0.64	0.68	1.4
2j8p	0.92	−0.72	0.81	1.4	0.88	−0.72	0.82	3.5
2ysb	0.88	0.25	0.81	170.6	0.87	0.25	0.76	11.6
2zaj	0.80	0.57	0.67	175.2	0.83	0.57	0.74	104.1
1w4h	0.93	0.89	0.91	88.7	0.93	0.89	0.74	6.3
1pgy	0.89	0.60	0.81	12.1	0.93	0.60	0.83	1.6

[a]The results are expressed in BCscore deviation to the experimental conformation. For each, we report the Protein Data Bank identifier (PDB), the BC-score of the best generated model (best), of the model of lowest energy (lowE), and the best model returned in the five best ranked models

dx.doi.org/10.1021/ct500592m | J. Chem. Theory Comput. 2014, 10, 4745−4758

Table 2. continued

according to the predicted TMscore (PEP-FOLD) or the Rosetta score (Rosetta). nEq: equivalent number of conformations generated (see Materials and Methods).

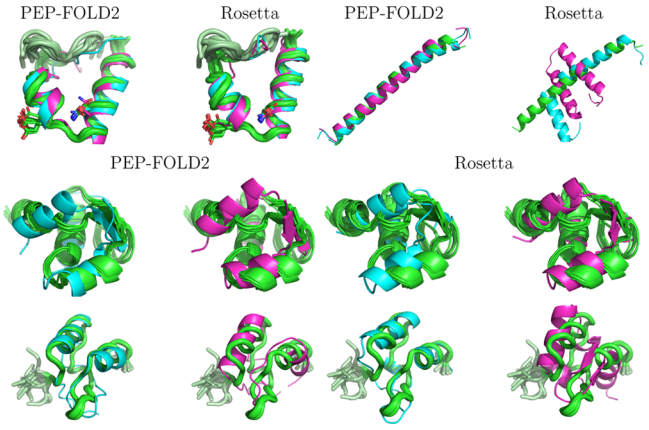


Figure 8. PEP-FOLD2 and Rosetta models. (green) Experimental conformation (pale green correspond to residues not in the rigid core). (cyan) Best model generated. (magenta) Best model returned in the five first ranks. (top) Models generated for 2l0g and 1usd. (middle, bottom) Models generated by PEP-FOLD2 and Rosetta for 1w4e and 1n9d.

an extended region and one helix, and 2k9d, a triple helix topology.

Finally, we have analyzed the diversity of the conformations generated by Rosetta and PEP-FOLD2. Table 2 reports the so-called equivalent number of conformations generated, neq (see eq 5), for each target. It is striking that, on average, the number of conformations explored by Rosetta is much smaller than that of PEP-FOLD—values of—4.48 and 43.52, respectively. For some targets such as 2ki0 and 1e0n, Rosetta generates conformations resulting in only one cluster, whereas PEP-FOLD2 generates 5.2 and 85.8 clusters, respectively, indicating that Rosetta introduces a much stronger bias in the conformational space sampling.

Figure 9A and B depicts the conformational diversity of the cluster centroid for 1wy3 using PEP-FOLD2 and Rosetta. Despite that fact that the nEq values are small and have similar magnitude, one notes the smaller difference in the conformations generated using Rosetta. Figure 9C and D depicts the neq information for 1wr3. Here, the difference between the neq values is large (24.7 versus 2.4). While only three clusters are generated using Rosetta, PEP-FOLD2 despite the bias levels used, explores a much larger conformational space, with even one of the conformations having a helical conformation instead of a strand one.

In our understanding, the reduced conformational diversity of conformations generated by Rosetta could result from the fact that Rosetta uses 9-mers fragments for 70% of the moves and then switches to 3-mer fragments for 30% of the moves,[10] while PEP-FOLD is exclusively based on 4-mers. If a stronger bias is of interest for modeling proteins with more than 80 amino acids, our results show that for peptides up to 50 amino acids, a larger conformational space can be considered at no loss of modeling performance.

CONCLUSIONS AND PERSPECTIVES

In this study, we have assessed the impact of revisiting the selection of the local conformations used by PEP-FOLD on the

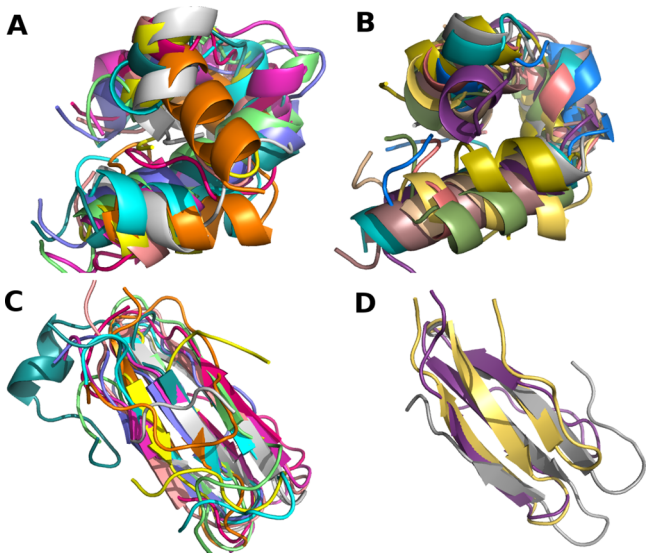


Figure 9. Diversity of the conformations returned by PEP-FOLD2 and Rosetta. Centroid of up to the ten best clusters when possible. A, B and C, D: models for 1wy3 and 1wr3 using PEP-FOLD2 and Rosetta, respectively.

structure prediction of 56 peptides with 25−52 amino acids. PEP-FOLD only attempts to identify the structure of lowest effective energy, and does not provide any information on how proteins fold. PEP-FOLD is available as free server and currently limited to the study of peptides with 9−37 amino-acids, but it cannot be downloaded yet. The PEP-FOLD server processes over several tens of thousands of requests a year since 2009 and has led to over 60 applications in the field of peptides or protein fragments (see ref 55), showing a need in the field of peptides and small proteins.

Each target was subject to 600 simulations using PEP-FOLD1, PEP-FOLD2, and Rosetta. Compared to PEP-FOLD1, PEP-FOLD2 improves significantly the quality of the models. PEP-FOLD1 fails for 4 targets and generates near-native models for 22 targets and native models for 26 targets. The corresponding values for PEP-FOLD2 are 3, 3, and 46 targets. PEP-FOLD2 is thus able to generate models consistent with experiments for 95% of the targets. Model recognition from the lowest sOPEP energy or even assisted by specialized tools such as Apollo remains to be improved. However, with no experimental structures at hand, PEP-FOLD2 proposes near-experimental models for 45 targets among 56 (80%) using the best five 5 ranked models, when PEP-FOLD1 could only propose such models for 38 targets.

Using the same criteria, Rosetta generates near-native or native conformations for 88% of the targets and proposes a near or experimental structure among the top five solutions for 75% of the targets. This 5% improvement in the performances of PEP-FOLD2 relative to Rosetta is non-negligible considering that PEP-FOLD2 explores a much larger conformational space than Rosetta, and the PEP-FOLD execution times are small, on the order of few minutes to few tens of minutes per simulation, depending on the size of the peptide. In their current implementations, PEP-FOLD2 runs 30% faster than Rosetta.

Finally, it is also interesting to note that despite the fact that Rosetta uses multiple coarse-grained representations and a final refinement with an all-atom model, PEP-FOLD2 with its single coarse-grained phase, generates approximately the same number of very high quality models (29 vs 34 sequence by Rosetta). This highlights the effectiveness of the sOPEP force field in its present version.

Being able to generate structural models compatible with experiments for peptide lengths of 25−53 amino acids at a high success rate, indicating that de novo peptide structure prediction by computers is approaching maturity, we are still not at the end of the game. Indeed, PEP-FOLD2 still faces several challenges. A first challenge is to understand why three targets among 56 cannot be generated correctly. Is it because the 2gdl and 2kya structures were solved by NMR in buffers containing 2.5 and 50% TFE and TFE effects are not treated by the sOPEP and Rosetta force field? For 2ovc, is it related to the fact it corresponds to a cytoplasmic fragment of a voltage-gated channel and thus might face nonstandard constraints? The force field sOPEP is based on the parametrization of the OPEP version 3.0 force field, but the accuray of OPEP has significantly improved from version 3 to version 5.[55] Whether the recognition of the native conformation from the lowest OPEP5 energy will improve remains to be determined. A second challenge is to determine whether the inclusion of the conformational entropy would help recognize the experimental structures among the top five or 10 models without resorting to very long simulations. We are exploring this aspect on several targets starting from the five best predicted models using simulated tempering.[56] A third one is to determine the relevance of the modeled conformations of the peptides for the study of peptide−protein complexes. Finally, another challenge is to keep moving toward high-throughput peptide analysis and to address previously unanswerable questions such as the screening of hundred thousands of peptide sequences per day using reasonable computer resources.

■ ASSOCIATED CONTENT

Supporting Information

Pictures of problematic targets using PEP-FOLD1 or Rosetta, as well as a picture of the best model generated for the 1yyb target using biased trajectories only. This material is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: pierre.tuffery@univ-paris-diderot.fr.

Present Addresses

[#]INSERM UMR-S 973, Univ Paris Diderot, Sorbonne Paris Cité, F-75205 Paris, France.

[‖]Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Institut de Biologie Physico-Chimique, F-75005 Paris, France.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank N. Mousseau and S. Coté for stimulating discussions about criteria to identify native-like models, and F. Guyon, for codeveloping the BCscore.

■ REFERENCES

(1) Vlieghe, P.; Lisowski, V.; Martinez, J.; Khrestchatisky, M. *Drug Discovery Today* **2010**, *15*, 40−56.
(2) Zhang, G.; Kazanietz, M. G.; Blumberg, P. M.; Hurley, J. H. *Cell* **1995**, *81*, 917−24.
(3) Mueller, T. D.; Feigon, J. *J. Mol. Biol.* **2002**, *319*, 1243−55.
(4) Alam, S. L.; Sun, J.; Payne, M.; Welch, B. D.; Blake, B. K.; Davis, D. R.; Meyer, H. H.; Emr, S. D.; Sundquist, W. I. *EMBO J.* **2004**, *23*, 1411−21.
(5) 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*, 235−42.
(6) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517−20.
(7) Zagrovic, B.; Snow, C. D.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 927−37.
(8) Escoubas, P.; King, G. F. *Expert Rev. Proteomics* **2009**, *6*, 221−4.
(9) Kim, D. E.; Chivian, D.; Baker, D. *Nucleic Acids Res.* **2004**, *32*, W526−31.
(10) Das, R.; Baker, D. *Annu. Rev. Biochem.* **2008**, *77*, 363−82.
(11) Zhang, Y. *BMC Bioinformatics* **2008**, *9*, 40.
(12) Xu, D.; Zhang, Y. *Proteins* **2012**, *80*, 1715−35.
(13) Kaur, H.; Garg, A.; Raghava, G. P. S. *Protein Pept. Lett.* **2007**, *14*, 626−31.
(14) Jayaram, B.; Bhushan, K.; Shenoy, S. R.; Narang, P.; Bose, S.; Agrawal, P.; Sahu, D.; Pandey, V. *Nucleic Acids Res.* **2006**, *34*, 6195−204.
(15) Thomas, A.; Deshayes, S.; Decaffmeyer, M.; Eyck, M.-H. V.; Charloteaux, B. B.; Brasseur, R. *Adv. Exp. Med. Biol.* **2009**, *611*, 459−60.
(16) Ozkan, S. B.; Wu, G. A.; Chodera, J. D.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 11987−92.
(17) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. *J. Phys. Chem. B* **2012**, *116*, 8494−503.
(18) Lee, J.; Lee, J.; Sasaki, T. N.; Sasai, M.; Seok, C.; Lee, J. *Proteins* **2011**, *79*, 2403−17.
(19) Nicosia, G.; Stracquadanio, G. *Biophys. J.* **2008**, *95*, 4988−4999.
(20) Chebaro, Y.; Pasquali, S.; Derreumaux, P. *J. Phys. Chem. B* **2012**, *116*, 8741−52.
(21) Chikenji, G.; Fujitsuka, Y.; Takada, S. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 3141−6.
(22) Voelz, V. A.; Shell, M. S.; Dill, K. A. *PLoS Comput. Biol.* **2009**, *5*, e1000281.
(23) Maupetit, J.; Derreumaux, P.; Tuffery, P. *Nucleic Acids Res.* **2009**, *37*, W498−503.
(24) Maupetit, J.; Derreumaux, P.; Tufféry, P. *J. Comput. Chem.* **2010**, *31*, 726−38.
(25) Camproux, A. C.; Gautier, R.; Tufféry, P. *J. Mol. Biol.* **2004**, *339*, 591−605.
(26) Thévenet, P.; Shen, Y.; Maupetit, J.; Guyon, F.; Derreumaux, P.; Tufféry, P. *Nucleic Acids Res.* **2012**, *40*, W288−93.
(27) Steckbeck, J. D.; Craigo, J. K.; Barnes, C. O.; Montelaro, R. C. *J. Biol. Chem.* **2011**, *286*, 27156−66.
(28) Feller, G.; Dehareng, D.; Lage, J.-L. D. *FEBS J.* **2011**, *278*, 2333−40.
(29) Berges, R.; Balzeau, J.; Takahashi, M.; Prevost, C.; Eyer, J. *PLoS One* **2012**, *7*, e49436.
(30) Olsson, N.; Wallin, S.; James, P.; Borrebaeck, C. A. K.; Wingren, C. *Protein Sci.* **2012**, *21*, 1897−910.
(31) López-Martínez, R.; Ramírez-Salinas, G. L.; Correa-Basurto, J.; Barrón, B. L. *PLoS One* **2013**, *8*, e76876.
(32) Gupta, S. K.; Singh, A.; Srivastava, M.; Gupta, S. K.; Akhoon, B. A. *Vaccine* **2009**, *28*, 120−31.
(33) Qureshi, A.; Thakur, N.; Tandon, H.; Kumar, M. *Nucleic Acids Res.* **2014**, *42*, D1147−53.
(34) Yan, L.; Yan, Y.; Liu, H.; LV, Q. *BioSystems* **2013**, *113*, 1−8.
(35) Wu, G.; Han, K.; LV, F. *J. Theor. Biol.* **2013**, *317*, 293−300.

(36) Horjales, S.; Schmidt-Arras, D.; Limardo, R. R.; Leclercq, O.; Obal, G.; Prina, E.; Turjanski, A. G.; Späth, G. F.; Buschiazzo, A. *Structure* **2012**, *20*, 1649−60.
(37) Francisco, B. S.; Bretsnyder, E. C.; Kranz, R. G. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, E788−97.
(38) Ulrich, E. L.; et al. *Nucleic Acids Res.* **2008**, *36*, D402−8.
(39) Hauser, M.; Mayer, C. E.; Söding, J. *BMC Bioinformatics* **2013**, *14*, 248.
(40) Wang, G.; Dunbrack, R. L. *Bioinformatics* **2003**, *19*, 1589−91.
(41) Viterbi, A. *IEEE Trans. Inform. Theory* **1967**, *13*, 260−269.
(42) Rabiner, L. *Proceedings of the IEEE* **1989**, *77*, 257−286.
(43) Tuffery, P.; Guyon, F.; Derreumaux, P. *J. Comput. Chem.* **2005**, *26*, 506−13.
(44) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389−402.
(45) Suzek, B. E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C. H. *Bioinformatics* **2007**, *23*, 1282−8.
(46) Guyon, F.; Camproux, A.-C.; Hochez, J.; Tufféry, P. *Nucleic Acids Res.* **2004**, *32*, W545−8.
(47) Maupetit, J.; Tuffery, P.; Derreumaux, P. *Proteins* **2007**, *69*, 394−408.
(48) Zemla, A.; Venclovas, C.; Moult, J.; Fidelis, K. *Proteins* **1999**, *Suppl 3*, 22−9.
(49) Zhang, Y.; Skolnick, J. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7594−9.
(50) Guyon, F.; Tufféry, P. *Bioinformatics* **2014**, *30*, 784−791.
(51) Wang, Z.; Eickholt, J.; Cheng, J. *Bioinformatics* **2011**, *27*, 1715−6.
(52) Labesse, G.; Colloc'h, N.; Pothier, J.; Mornon, J. P. *Comput. Appl. Biosci.* **1997**, *13*, 291−5.
(53) Mirabello, C.; Pollastri, G. *Bioinformatics* **2013**, *29*, 2056−8.
(54) Buchan, D. W.; Minneci, F.; Nugent, T. C.; Bryson, K.; Jones, D. T. *Nucleic acids research* **2013**, *41*, W349−W357.
(55) Sterpone, F.; # 0Melchionna, S.; Tuffery, P.; Pasquali, S.; Mousseau, N.; Cragnolini, T.; Chebaro, Y.; St-Pierre, J.-F.; Kalimeri, M.; Barducci, A.; Laurin, Y.; Tek, A.; Baaden, M.; Nguyen, P. H.; Derreumaux, P. *Chem. Soc. Rev.* **2014**, *43*, 4871−93.
(56) Nguyen, P. H.; Okamoto, Y.; Derreumaux, P. *J. Chem. Phys.* **2013**, *138*, 061102.
(57) Rey, J.; Tufféry, P. *BactPepDB: a database of predicted peptides from a exhaustive survey of complete prokaryote genomes*, submitted.