PEP-FOLD: an online resource for de novo peptide structure prediction

Julien Maupetit¹,², Philippe Derreumaux³ and Pierre Tuffery¹,²,*

¹MTI, INSERM UMR-S 973, - Paris 7, 35 rue H. Brion, F75205, Paris, ²RPBS, Université Paris Diderot - Paris 7, 35 rue H. Brion, F75205, Paris and ³Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, F75005, Paris, France

Received January 30, 2009; Revised April 5, 2009; Accepted April 20, 2009

ABSTRACT

Rational peptide design and large-scale prediction of peptide structure from sequence remain a challenge for chemical biologists. We present PEP-FOLD, an online service, aimed at de novo modelling of 3D conformations for peptides between 9 and 25 amino acids in aqueous solution. Using a hidden Markov model-derived structural alphabet (SA) of 27 four-residue letters, PEP-FOLD first predicts the SA letter profiles from the amino acid sequence and then assembles the predicted fragments by a greedy procedure driven by a modified version of the OPEP coarse-grained force field. Starting from an amino acid sequence, PEP-FOLD performs series of 50 simulations and returns the most representative conformations identified in terms of energy and population. Using a benchmark of 25 peptides with 9–23 amino acids, and considering the reproducibility of the runs, we find that, on average, PEP-FOLD locates lowest energy conformations differing by 2.6 Å Cα root mean square deviation from the full NMR structures. PEP-FOLD can be accessed at http://bioserv.rpbs.univ-paris-diderot.fr/PEP-FOLD

INTRODUCTION

While computational biologists and chemists have developed convincing ab initio or de novo approaches for protein structure determination [see CASP experiments (1)], accurate and fast peptide 3D conformation prediction from bioinformatics-based or physics-based methods is still an open challenge. Yet, peptides play many biological functions ranging from hormones, neurotransmitters to antibiotics, among others (2,3).

The obstacles for elucidating the relationship between peptide sequences, structures and functions arise from several factors. First, in contrast to proteins, short peptides do not systematically adopt a stable well-defined tertiary structures (4). Then, the experimental flow of peptide structure determination based on Nuclear Magnetic Resonance (NMR) spectroscopy and X-ray crystallography remains very low, and structures can also be determined by using non-physiological conditions to increase their solubility such as trifluoroethanol (TFE)/water mixtures. In addition, many peptides are marginally stable, and early peptide structure determinations may be questioned. For instance, the 20-residue betanova peptide was initially proposed as adopting a three-stranded β-structure in aqueous solution (5), but its content of β-sheet was revisited and reduced by long molecular dynamics (MD) simulations and new experiments to 10% (6). Finally, on the theoretical front, in spite of constant effort, large-scale in silico prediction using all-atom models is still not feasible. The application of the Rosetta approach to peptides remains to be evaluated (7), and all-atom molecular dynamics or Monte Carlo simulations in explicit or implicit solvent models with replica exchanges are hampered by the computer time and resources (8,9).

To accelerate conformational search and reduce the number of local minima, many studies have resorted to simplified representations and energy models, but the main problem is to preserve the physics of the systems (10). In this context, Ichikawa et al. (11) proposed a growing chain algorithm using a number of discrete dihedral \( \phi/\psi \) choices and a sum of hydrophobic and hydrogen bond interactions. Raghava et al. (12) designed PepStr based on secondary structure and \( \beta \)-turn prediction with a short MD-based energy refinement. For their part, Thomas et al. (13) developed Peplook based on a Boltzmann-stochastic algorithm coupled to 64 \( \phi/\psi \) backbone combinations, but their method failed to predict all topologies. More recently, Nicosia and Stracquadanio (14) proposed a generalized pattern search algorithm (GPS) using an all-atom energy model and secondary structure prediction (for systems with more 15 amino acids). Using a benchmark of 42 peptides with 9–20 amino acids in aqueous and non-aqueous solutions, GPS approached the experimental structures at 3.2 Å Cα root mean square deviation.
(cRMSD) (14). Currently, the best server available to biologists for peptide structure prediction is the PepStr server reaching an averaged cRMSD accuracy of 4.0 Å on a set of 42 peptides (12).

Here we present, the PEP-FOLD server, which builds on a new de novo approach to predict 3D peptide structures from sequence information. PEP-FOLD is based on the concept of structural alphabet (SA) and uses a Hidden Markov Model (HMM)-derived SA of 27 letters to describe proteins as series of overlapping fragments of four amino acids (15). PEP-FOLD uses a two-step procedure: prediction of a limited set of SA letters at each position from sequence, and then assembly of the prototype fragments associated with each SA letter using a revised version of our greedy algorithm (16,17) and a generic protein coarse-grained force field (18).

PEP-FOLD is very fast and can be used with high confidence for any natural peptide sequence of 9–25 amino acids in aqueous solution, and, with more caution, in non-aqueous solutions. All details about the algorithm, force field and the results can be found elsewhere (Maupetit, J. et al., submitted for publication).

CONCEPTS AND METHODS

Figure 1 presents a general overview of PEP-FOLD.

SA profile prediction from amino acid sequence

PEP-FOLD uses a HMM-derived SA of 27 letters, describing proteins as series of overlapping fragments of four amino acids (15). The conformations of consecutive fragments are not independent and follow a first-order Markovian process. An SA letter at a position cannot therefore be followed by all 27 SA letters at the adjacent position. A protein of \( L \) amino acids is represented by a SA sequence of \( L - 3 \) SA letters. To learn SA letter profiles from the amino acid sequence, PEP-FOLD uses the following two-step procedure.

In the first step, the SA letter profiles are predicted at each independent position, using an amino acid profile obtained by running PSI-BLAST (19) against the uniref data-bank (20). As support vector machine (SVM) input, we use the amino acid profile of each four-residue fragment enlarged by two residues on both sides, and therefore a \( 20 \times 8 \) dimensional vector to predict one SA letter. The SVM output is a profile of \( L - 3 \times 27 \) probabilities that gives the predicted probability that each SA letter fits each fragment of the protein. We emphasize that this prediction step is free of any bias. The SVM training was performed by using a collection of structures excluding our peptide set.

In the second step, these probabilities are used as input data of a forward-backward algorithm along with the HMM-SA model. We have observed that the prediction of only the eight best letters at each position contains the native conformation information. Thus, the output of the SA prediction is a profile of \( L - 3 \times 8 \) SA letters.

Note that the PEP-FOLD server also allows the user to constrain the predicted SA profile by PSIPRED secondary structure prediction (21). This is done, for the amino acid positions predicted by PSIPRED as helices (or \( \beta \)-strands) with a confidence index >5, by restraining their possible SA letters to helical (or \( \beta \)-strand) letters. Although we found that the PSIPRED option has a marginal impact in most cases, it can sometimes help identify the native conformation.

3D assembly of prototype conformations

To build 3D structures from the predicted SA letter profiles, we perform a rigid assembly of the predicted fragments by using an enhanced version of our greedy algorithm (16,17). Whereas the early algorithm uses forward (from N- to C-terminal) and backward (from C- to N-terminal) incremental operators to grow the chain, our revisited version uses a zip operator to start the building process at any position of the structure, alternatively adding \( S \) residues at each side of the growing structure. After extensive tests on 9–25 residue systems, we found that an \( S \) value of 1 is appropriate, the starting position can be selected randomly, and very high accuracy can be achieved by using a total of 74 prototypes to describe the 27 SA letters. 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.

In contrast to our previous studies where the assembly was guided by an energy function dependent on the knowledge of the native long-range $C_\alpha-C_\alpha$ contacts (16,17), PEP-FOLD assembly is driven by the generic protein coarse-grained force field, OPEP 3.1 version (18), slightly adapted for greedy algorithm. As many force fields, OPEP includes local interactions for stereochemistry (e.g. bond lengths, bond angles, etc.), van der Waals interactions between all particles and terms allowing the formation of hydrogen bonds. OPEP uses the positions of the backbone N, H, C\(\alpha\), C and O atoms and one bead for each side chain, and is designed for basin hopin (10) and molecular dynamics (22) simulations in Cartesian coordinate space. Since the greedy algorithm works in a discrete space and uses well-defined fragments of four-residues, we derived sOPEP. The main difference between OPEP and sOPEP comes from the formulation of the side chain—side chain interactions that result in smoother interactions at short distances, reducing therefore the number of steric clashes during rigid assembly.

3D models post-treatment

Each polypeptide is subject to one run, i.e. 50 greedy simulations differing in the starting position (see previous section) and randomness in the retained candidates at each position. The conformational diversity of the 50 models is characterized by cluster analysis using a complete linkage procedure and a threshold of 2 Å cRMSD for peptides less than 20 amino acids and 3 Å above. For each cluster, we return the lowest energy centroid conformation.

The Greedy-OPEP assembling procedure generates coarse-grained structure models. These models are post-treated in two ways. First, the side chain beads are discarded and all-atom side-chains are positioned using a fast backbone-dependent procedure previously used in the SABBAC server (23). Secondly, to provide high quality structures, a fast minimization is performed with GROMACS (24).

INPUT/OUTPUT

PEP-FOLD accepts, as input, sequences of lengths between 9 and 25 amino acids. Most of the frequent sequence formats including the FASTA or RAW formats are accepted. Only standard amino acids are presently accepted, and no hetero groups—including ions—are considered. Disulfide bonds are not managed. It is possible to constrain the SA local structure prediction using PSIPRED, although this is not active by default. Finally, for the sake of convenience, the user can specify a reference structure for comparison with the de novo models.

On program termination, the server gives the results of the cluster analysis. Although the number of clusters, Ncs, can vary from 1 (all conformations similar) to 50, Ncs is usually less than 5. The output first returns a 2D plot with the population and energy of each cluster and the localization of the lowest energy (Figure 2A), and then a picture of the centroid of each cluster (Figure 2B). Subsequently, the output gives for each of the 50 conformations its cluster assignment and energy. If a reference structure was specified, the cRMSD, the GDT-TS (25) and TM scores (26) of each model are also reported. Finally, an archive containing all cluster centroid conformations and the lowest energy conformation (LEC) (PDB format) is available for download. The LEC is also available separately.

Since PEP-FOLD is based on a chain growth algorithm, the CPU times vary linearly with peptide length. Typical execution times, using 10 Intel Xeon 3GHz processing units in parallel, vary for one run from 10 min for a 10-residue peptide to 45 min for a 25-residue peptide.

PERFORMANCES

In a previous report, the PEP-FOLD performances were assessed using a benchmark of 25 peptides with 9–25
LEC identity: the variation in energy maximal energy difference (Max $E$) between the five LECs from runs 1–5. We also give the minimal and maximal cRMSD (in angstroms) of the five LECs with respect to the NMR structure (native cRMSD). Finally, we report the average cRMSD between the five LECs (LEC cRMSD). Note that 1pgbF corresponds to the fragment 41–56 of 1pgb, and the NMR structure used for cRMSD calculation is model 1, except for 1a13, 1b03, 1e0q, legs, 1lcx, 1pei, 1rpv, and 1wz4 where the reference models are 10, 4, 8, 3, 25, 10, 12, 23 and 2, respectively.

In summary, using our strict criteria for reproducibility, the probability to locate the LEC by using one run of 50 greedy simulations for a peptide with 9–25 amino acids is 86%. The probability reaches 96% in terms of convergence towards the same topology.

We present in Figure 3 the results on the previously studied 10-residue peptide chignolin (PDB code: 1UAO) (27), and a new system, the 23-residue magainin antibiotic peptide (PDB code: 2MAG) (3).

For the chignolin, the 18 NMR models are very similar and display a $\beta$-hairpin. PEP-FOLD generates four distinct conformations. The lowest energy model displays a fully native structure (cRMSD of 2.0 Å) with the experimental $\beta$-strands spanning residues 2–3/8–9 correctly predicted. The $\beta$-carbon RMSD is 1.0 Å from the NMR structure, indicating correct orientation of the side chains relative to the strands as seen in Figure 3.

The magainin antibiotic peptide is soluble, but unstructured in aqueous solution. Its NMR structure in dipalmitylophosphatidylcholin (DPC), sodium dodecyl sulfate (SDS) and TFE/water environments is characterized by a long $\alpha$-helix that is kinked at residues Phe12-Gly13 (3). The conformation of the first cluster identified by PEP-FOLD is very similar to the NMR structure in non-aqueous solution (0.7 Å cRMSD, see the green and magenta structures in Figure 3) and correctly predicts the helix kink. The second cluster of lowest energy energy, only 2 kcal/mole more stable than the first cluster, displays two $\alpha$-helices spanning 1–12 and 17–22 stabilized by an hydrophobic patch between the three phenylalanines (Phe5, Phe12 and Phe16). Compared to the

### Table 1. PEP-FOLD results on 25 peptides with 9–25 amino acids in aqueous solution

<table>
<thead>
<tr>
<th>PDB</th>
<th>$L$</th>
<th>Class</th>
<th>LEC energy</th>
<th>Max $E$</th>
<th>Native cRMSD</th>
<th>LEC cRMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a13</td>
<td>14</td>
<td>–</td>
<td>-26.5</td>
<td>0.0</td>
<td>1.8</td>
<td>0.0</td>
</tr>
<tr>
<td>1b03</td>
<td>18</td>
<td>B</td>
<td>-23.6</td>
<td>1.4</td>
<td>2.0–3.1</td>
<td>2.1</td>
</tr>
<tr>
<td>1dep</td>
<td>15</td>
<td>A</td>
<td>-30.7</td>
<td>0.0</td>
<td>1.7</td>
<td>0.0</td>
</tr>
<tr>
<td>1du1</td>
<td>20</td>
<td>A</td>
<td>-45.1</td>
<td>0.0</td>
<td>5.1</td>
<td>0.0</td>
</tr>
<tr>
<td>1e0q</td>
<td>17</td>
<td>B</td>
<td>-26.7</td>
<td>0.2</td>
<td>2.7–5.7</td>
<td>2.2</td>
</tr>
<tr>
<td>legs</td>
<td>9</td>
<td>–</td>
<td>-3.9</td>
<td>0.0</td>
<td>1.5</td>
<td>0.0</td>
</tr>
<tr>
<td>1gf</td>
<td>14</td>
<td>A</td>
<td>-24.7</td>
<td>0.0</td>
<td>2.4–2.5</td>
<td>0.1</td>
</tr>
<tr>
<td>lin3</td>
<td>12</td>
<td>A</td>
<td>-25.8</td>
<td>0.0</td>
<td>2.3</td>
<td>0.0</td>
</tr>
<tr>
<td>1k43</td>
<td>14</td>
<td>B</td>
<td>-20.9</td>
<td>0.4</td>
<td>1.5–1.7</td>
<td>0.8</td>
</tr>
<tr>
<td>1l2y</td>
<td>20</td>
<td>A</td>
<td>-28.5</td>
<td>0.1</td>
<td>2.1–2.3</td>
<td>0.2</td>
</tr>
<tr>
<td>1l3q</td>
<td>12</td>
<td>–</td>
<td>-6.1</td>
<td>0.7</td>
<td>3.3–5.3</td>
<td>2.5</td>
</tr>
<tr>
<td>1lcx</td>
<td>13</td>
<td>–</td>
<td>-28.0</td>
<td>0.3</td>
<td>2.8</td>
<td>0.0</td>
</tr>
<tr>
<td>1le1</td>
<td>12</td>
<td>B</td>
<td>-18.2</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1le3</td>
<td>16</td>
<td>B</td>
<td>-24.2</td>
<td>3.9</td>
<td>1.2–2.7</td>
<td>1.1</td>
</tr>
<tr>
<td>1niz</td>
<td>14</td>
<td>B</td>
<td>-19.1</td>
<td>0.7</td>
<td>2.1–2.7</td>
<td>0.7</td>
</tr>
<tr>
<td>1nkf</td>
<td>16</td>
<td>A</td>
<td>-18.9</td>
<td>0.0</td>
<td>4.3–4.5</td>
<td>0.3</td>
</tr>
<tr>
<td>1pef</td>
<td>18</td>
<td>A</td>
<td>-58.2</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
</tr>
<tr>
<td>1pe1</td>
<td>22</td>
<td>A</td>
<td>-50.3</td>
<td>0.0</td>
<td>1.6</td>
<td>0.0</td>
</tr>
<tr>
<td>1pgbF</td>
<td>16</td>
<td>B</td>
<td>-21.6</td>
<td>2.7</td>
<td>2.1–2.4</td>
<td>2.4</td>
</tr>
<tr>
<td>1rpv</td>
<td>17</td>
<td>A</td>
<td>-28.9</td>
<td>0.0</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>1ua0</td>
<td>10</td>
<td>B</td>
<td>-7.3</td>
<td>0.1</td>
<td>2.0</td>
<td>0.1</td>
</tr>
<tr>
<td>1wbr</td>
<td>17</td>
<td>–</td>
<td>-33.9</td>
<td>0.0</td>
<td>3.5</td>
<td>0.0</td>
</tr>
<tr>
<td>1wz4</td>
<td>23</td>
<td>A</td>
<td>-31.2</td>
<td>0.6</td>
<td>5.6–6.4</td>
<td>2.8</td>
</tr>
<tr>
<td>2nl</td>
<td>15</td>
<td>–</td>
<td>-23.5</td>
<td>0.0</td>
<td>4.5</td>
<td>0.0</td>
</tr>
<tr>
<td>2eqv</td>
<td>12</td>
<td>B</td>
<td>-16.9</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
</tr>
</tbody>
</table>

PDB: Protein Data Bank identifier. $L$: peptide length. Class: structural class (A for $\alpha$, B for $\beta$ and - for none), according to STRIDE (29) program. Each target is subject to five independent runs, each of 50 greedy simulations. For each peptide, we report the LEC (kcal/mole) from run 1 and the maximal energy difference (Max $E$) between the five LECs from runs 1–5. We also give the minimal and maximal cRMSD (in angstroms) of the five LECs with respect to the NMR structure (native cRMSD). Finally, we report the average cRMSD between the five LECs (LEC cRMSD). Note that 1pgbF corresponds to the fragment 41–56 of 1pgb, and the NMR structure used for cRMSD calculation is model 1, except for 1a13, 1b03, 1e0q, legs, 1lcx, 1pei, 1rpv, 1wbr and 1wz4 where the reference models are 10, 4, 8, 3, 25, 10, 12, 23 and 2, respectively.

Table 1 summarizes the performances without any PSIPRED constraints, using five independent runs. It reports for each peptide the LEC obtained from run 1, the maximal energy difference between the five LECs, the extremal cRMSD between the five LECs and the NMR structure, and the average cRMSD between the five LECs. Averaged on 25 peptides and five runs, the PEP-FOLD LEC reproduces the NMR structure at 2.6 Å cRMSD. Considering a strict criterion to accept LEC identity: the variation in energy $< 1$ kcal/mole and variation in cRMSD $< 1.2$ Å, we find that for 20 peptides, the LEC remains the same between the 5 runs, and the average cRMSD between the LECs is of 0.6 Å. For 1le3, 1b03 1pgbF, the energy differences between the LECs do not change the predicted topology. The maximal energy difference of 1.4 kcal/mole with a cRMSD difference of 1.1 Å leads, however, for 1b03, to a $\beta$-hairpin with out-of-register strands. In contrast, for 1l3q and 1e0q, the NMR topology is not identified systematically: for 1e0q, a LEC of -27 kcal/mole displays the NMR $\beta$-hairpin topology, but PEP-FOLD frequently returns a LEC with helix-turn-strand character.

In summary, using our strict criteria for reproducibility, the probability to locate the LEC by using one run of 50 greedy simulations for a peptide with 9–25 amino acids is 86%. The probability reaches 96% in terms of convergence towards the same topology.

We present in Figure 3 the results on the previously studied 10-residue peptide chignolin (PDB code: 1UAO) (27), and a new system, the 23-residue magainin antibiotic peptide (PDB code: 2MAG) (3).

For the chignolin, the 18 NMR models are very similar and display a $\beta$-hairpin. PEP-FOLD generates four distinct conformations. The lowest energy model displays a fully native structure (cRMSD of 2.0 Å) with the experimental $\beta$-strands spanning residues 2–3/8–9 correctly predicted. The $\beta$-carbon RMSD is 1.0 Å from the NMR structure, indicating correct orientation of the side chains relative to the strands as seen in Figure 3.

The magainin antibiotic peptide is soluble, but unstructured in aqueous solution. Its NMR structure in dipalmitylophosphatidylcholin (DPC), sodium dodecyl sulfate (SDS) and TFE/water environments is characterized by a long $\alpha$-helix that is kinked at residues Phe12-Gly13 (3). The conformation of the first cluster identified by PEP-FOLD is very similar to the NMR structure in non-aqueous solution (0.7 Å cRMSD, see the green and magenta structures in Figure 3) and correctly predicts the helix kink. The second cluster of lowest energy energy, only 2 kcal/mole more stable than the first cluster, displays two $\alpha$-helices spanning 1–12 and 17–22 stabilized by an hydrophobic patch between the three phenylalanines (Phe5, Phe12 and Phe16). Compared to the
peptides in non-aqueous solutions (2.7 Å cRMSD) (results not shown), we are currently optimizing sOPEP parameters for these environments. Another direction we are exploring is the possibility to treat linear and cyclic peptides combining both D- and L-amino acids, and peptides with disulfide bridges. Overall, we believe that PEP-FOLD, with its present accuracy and speed, should be a useful server for chemical biologists.

**FUNDING**

Funding for open access charge: INSERM UMR-S 973.

Conflict of interest statement. None declared.

**REFERENCES**


