Unfolding the fold of cyclic cysteine-rich peptides

AMARDA SHEHU,[1] LYDIA E. KAVRAKI,[1,2,3] AND CECILIA CLEMENTI[2,4]

[1]Department of Computer Science, Rice University, Houston, Texas 77005, USA
[2]Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas 77030, USA
[3]Department of Bioengineering, Rice University, Houston, Texas 77005, USA
[4]Department of Chemistry, Rice University, Houston, Texas 77005, USA

(RECEIVED July 25, 2007; FINAL REVISION November 2, 2007; ACCEPTED December 14, 2007)

Abstract

We propose a method to extensively characterize the native state ensemble of cyclic cysteine-rich peptides. The method uses minimal information, namely, amino acid sequence and cyclization, as a topological feature that characterizes the native state. The method does not assume a specific disulfide bond pairing for cysteines and allows the possibility of unpaired cysteines. A detailed view of the conformational space relevant for the native state is obtained through a hierarchic multi-resolution exploration. A crucial feature of the exploration is a geometric approach that efficiently generates a large number of distinct cyclic conformations independently of one another. A spatial and energetic analysis of the generated conformations associates a free-energy landscape to the explored conformational space. Application to three long cyclic peptides of different folds shows that the conformational ensembles and cysteine arrangements associated with free energy minima are fully consistent with available experimental data. The results provide a detailed analysis of the native state features of cyclic peptides that can be further tested in experiment.

Keywords: native state ensemble; free-energy landscape; cysteine rearrangements; cyclic cysteine-rich peptides

Supplemental material: see www.proteinscience.org

Chemical and physical studies on cysteine-rich enzymes led Anfinsen (1973) to postulate that the amino acid (aa) sequence encodes for the correct tertiary structure and arrangement of cysteine residues in disulfide bonds in the protein native state. Since then, experiment, computation, and theory have shown functional relevance both in excursions of a protein from an average experimentally determined native structure (Schnell et al. 2004; Eisenmesser et al. 2005; Karplus and Kuriyan 2005) and in rearrangements of cysteines in different disulfide bonds under native conditions (Hogg 2003). Experimental and computational characterization of the structural diversity of the native state and the diversity of cysteine arrangements remain active areas of research (Hilser et al. 1998; Palmer et al. 2001; Czaplewski et al. 2004; Lindorff-Larsen et al. 2005; Shehu et al. 2006).

In this work, we propose a computational method to characterize the native state of cyclic cysteine-rich peptides. In the following, we refer to this method as NcCYP for native state characterization of cysteine-rich cyclic peptides. NcCYP uses minimal information—more specifically (1) amino acid sequence and (2) backbone cyclization—to generate low-energy conformations comprising the native state. No a priori assumptions are made about the native disulfide bond pairing of cysteines. Proximity and energetic criteria determine how to feasibly arrange cysteines in each conformation generated, also allowing the possibility of unpaired cysteines.

Reprint requests to: Cecilia Clementi, Department of Chemistry, Rice University, Houston, Texas 77005, USA; e-mail: cecilia@rice.edu; fax: (713) 348-5155; or Lydia E. Kavraki, Department of Bioengineering, Rice University, Houston, Texas 77005, USA; e-mail: kavraki@rice.edu; fax: (713) 348-5930.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.073142708.

Many methods (mostly based on Molecular Dynamics or Monte Carlo procedures) have been proposed to target cyclic peptides (Krause et al. 2000; Loiseau et al. 2003; Rayan et al. 2004) as these peptides' enhanced stability and diverse biological activities are appealing for peptidomimetics and pharmaceutical purposes (Craik 2006; Craik et al. 2006; Shin et al. 2007). NcCYP presents two improvements over current computational methods: (1) an efficient exploration of the conformational space allows us to generate a very large number (hundreds of thousands) of low-energy cyclic conformations in reasonable time, and (2) the diversity of cysteine arrangements is considered.

In practice, this is achieved first by obtaining conformations satisfying the geometric constraints imposed by cyclization and then subjecting these conformations to energetic refinement. This two-step procedure is necessary, since geometrical considerations alone do not guarantee low-energy conformations. Such treatment has been shown both general and efficient in generating large ensembles of native conformations for proteins (Shehu et al. 2006, 2007a,b).

According to the protein energy landscape perspective (Onuchic et al. 1997), native conformations are associated with the global minimum on a smooth free-energy landscape. The high-dimensionality of the protein conformational space poses significant demands on computational methods searching for the global minimum on this landscape. Current computational methods can perform this search using additional information about the native state, in the form of experimentally obtained data (Lindorff-Larsen et al. 2005) or average native structures (Shehu et al. 2006, 2007a). NcCYP uses a hierarchic multi-resolution exploration to efficiently explore the high-dimensional space of conformations relevant for the native state in cyclic peptides without using additional information of experimental data. While the conformational space remains vast (the peptides considered in this work are up to 31 aa long), compared with proteins, it is more tractable to exploration.

Methods that search for native conformations traditionally do not allow for cysteine rearrangements in different disulfide bonds, even though experiments show that cysteine rearrangements may drive function, misfolding, or disease (Hogg 2003; Baneyx and Mujacic 2004). The prediction of cysteine arrangements has been addressed previously by using statistical mechanics (Fiser et al. 1992), optimized threading potentials (Dombkowski and Crippen 2000), neural networks (Martelli et al. 2004; Ferré and Clote 2005), or sequence information (Mucchielli-Giorgi et al. 2002). Usually, methods focused on generating conformations in the context of protein structure prediction or protein folding prefer to treat disulfide bonds as fixed constraints during the course of the simulation to reduce the dimensionality of the search space (Skolnick et al. 1997; Abkevich and Shakhnovich 2000). Recent attempts to allow cysteine rearrangements during the search for native conformations often result in low-energy conformations with nonnative arrangements (Czaplewski et al. 2004).

NcCYP does not assume a specific disulfide bond pairing between cysteines; neither do all cysteines need to be paired for a resulting conformation to be energetically feasible a priori. In the following, we briefly summarize related work and the main ingredients of the proposed method before discussing the results.

Summary of the proposed method

NcCYP generates conformations through a multi-resolution approach. Cyclic backbones are first obtained employing a low-resolution representation that considers only backbone atoms. Each cyclic backbone is converted to an all-atom conformation, which is then energetically refined. A physically realistic all-atom force field in implicit solvent is used to associate an energetically favorable cysteine arrangement to each conformation (for details, see Materials and Methods).

This approach allows us to generate a large number of all-atom conformations with distinct cyclic structures and feasible cysteine arrangements. Conformations are clustered according to their cysteine arrangements to reveal low-energy minima associated with different arrangements. Conformations representative of energy minima are selected and used as starting points in the search for new lower-energy conformations. This iterative exploration continues until no lower-energy minima are obtained.

Generated conformations are analyzed through a spatial and energetic analysis. Nonlinear dimensionality reduction is employed to reveal global reaction coordinates that structurally distinguish among conformations. These coordinates allow us to visualize conformational clusters and associate a free-energy landscape to the explored space. Comparing free energies of emerging clusters yields a probability distribution for the possible cysteine arrangements.

Results

We present the results obtained from applications of NcCYP to three cyclic cysteine-rich peptides of different lengths and folds. Two peptides are naturally occurring; the other is cyclized from a naturally occurring peptide.

Rhesus θ-defensin-1

The first system selected for our study is rhesus θ-defensin-1 (RTD-1), a cyclic peptide found in Rhesus macaque leukocytes. As part of the immune system (Tang

et al. 1999), RTD-1 is microbicidal and three times more potent than its open-chain human analog (Tang et al. 1999). RTD-1 consists of 18 aa and assumes a β-hairpin fold under native conditions (Trabi et al. 2001; N.L. Daly, Y.-K. Chen, K.J. Rosengren, U.C. Marx, M.L. Phillips, A.J. Waring, W. Wang, R.I. Lehrer, and D.J. Craik, in prep.). The NMR ensemble (Protein Data Bank [PDB] code 2atg) (Trabi et al. 2001; N.L. Daly, Y.-K. Chen, K.J. Rosengren, U.C. Marx, M.L. Phillips, A.J. Waring, W. Wang, R.I. Lehrer, and D.J. Craik, in prep.) in Figure 1A shows flexible turns connecting the β-sheets. The three disulfide bonds in this ensemble are in a ladder arrangement between cysteines 4–17, 6–15, and 8–13.

cMII-6

The second system selected for the application of NcCYP is cyclized by adding a 6-aa linker to the naturally occurring α-conotoxin (α-CTX) MII (Clark et al. 2005). Found in the venom of *Conus magus*, MII is a potent inhibitor of nicotinic acetylcholine receptors and a potential drug lead against Parkinson's disease (Quik et al. 2001). The MII NMR ensemble (PDB code 1MII) and 16-aa sequence are shown in Figure 1B. Cyclization of MII is possible because the 11.2 ± 0.3 Å distance between the termini can be spanned by a few amino acids (Clark et al. 2005). The sequence of the linker and the resulting cMII-6 are given in Figure 1C. The cMII-6 NMR ensemble (PDB code 2AJW) (Clark et al. 2005) in Figure 1D shows that, while the linker is highly flexible (as expected from its richness in GLY and ALA), cMII-6 retains both MII's central α-helix and the two disulfide bonds between cysteines 8–14 and 9–22.

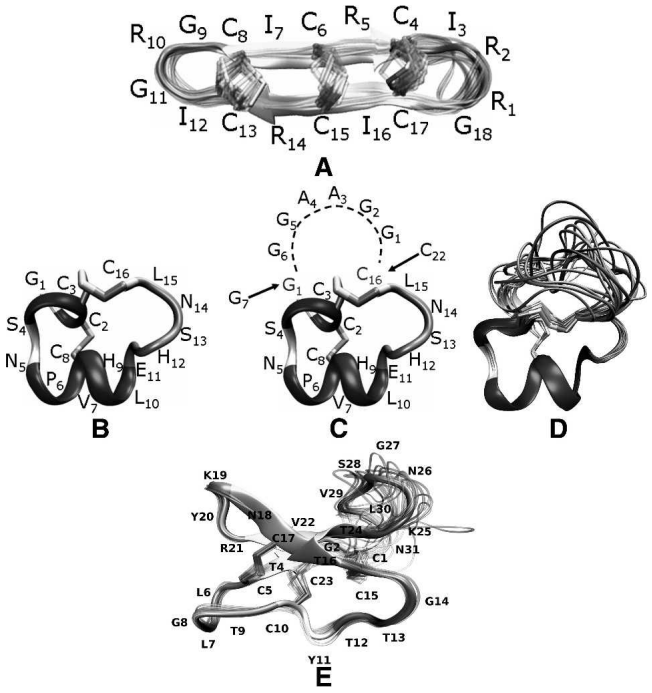

Figure 1. All 20 conformations corresponding to NMR ensembles of RTD-1, MII, cMII-6, and kalata B8 are shown superimposed over one another in *A*, *B*, *D*, and *E*, respectively, with the first conformation of each ensemble drawn thicker for reference. Sequences are shown for each peptide. (*C*) The dashed line delineates the 6-aa sequence of the linker in cMII-6, which shifts amino acid positions by six.

Kalata B8

The final system selected for this study is kalata B8, a cyclic peptide found in *Oldenlandia affinis*. Kalata B8 is a hybrid of the two major Möbius and bracelet subfamilies of the cyclotide family (Daly et al. 2006). Like other cyclotides, kalata B8 displays anti-HIV activity. Unlike other cyclotides, the peptide exhibits significant conformational flexibility in the native state while maintaining its cysteine knot motif (Daly et al. 2006). Figure 1E shows the 20 structures of the NMR ensemble of the 31-aa chain of kalata B8 (PDB code 2b38), with the sequence superimposed over the ensemble. Figure 1E also shows the six cysteines paired up in the 1–10, 5–17, and 10–23 arrangement.

Generation of conformational ensembles

We apply NcCYP, described in detail in the Materials and Methods section, to the RTD-1, cMII-6, and kalata B8 sequences to generate a large number of low-energy all-atom conformations, 534,299, 518,100, and 539,205, respectively. These conformations provide a broad view of the space relevant for the native state. Analysis of the associated energy landscape is performed by considering generated conformations with potential energies no higher than 20 kcal/mol from the global minimum potential energy obtained for each peptide. This choice is justified by the fact that conformations with higher energy values have a negligible Boltzmann probability ($\leq 10^{-15}$) at room temperature.

The 8034, 5380, and 4420 RTD-1, cMII-6, and kalata B8 conformations, respectively, that satisfy this energetic criterion are projected onto a lower-dimensional space through Scalable Isomap (ScIMAP), a scalable nonlinear dimensionality reduction technique (Das et al. 2006) developed by our groups. ScIMAP first selects several conformations to serve as ''landmarks.'' The nearest-neighbors graph, which connects each conformation to its nearest neighbors according to least root-mean-squared-deviation (lRMSD), is used to compute distances from landmarks to the conformations in the ensemble. A low-dimensional projection is then obtained by using the top eigenvectors from the resulting distance matrix as a base set. Previous work has shown that ScIMAP can

accurately reveal the few essential coordinates that capture the variability of conformations associated with large molecular motions (Das et al. 2006).

The application of ScIMAP in this work shows that two coordinates are sufficient to capture more than 90% of the structural variability among the conformations obtained for each peptide. By use of a modified version of the weighted histogram method (WHAM) (Ferrenberg and Swendsen 1988, 1989), free energy calculations on the low-dimensional landscapes identify the cysteine arrangements that are most probable at room temperature. Finally, conformations associated with lowest free energy states are subjected to explicit solvent equilibrations (for details, see the Supplemental material) to ensure that structural stability is maintained across different solvation models.

Two main results emerge from the analysis summarized below: (1) on both naturally occurring RTD-1 and kalata B8 peptides, the native cysteine arrangement present in the NMR ensemble is correctly recovered as the lowest free energy state; and (2) on the engineered cMII-6 peptide, NcCYP predicts that two distinct cysteine arrangements can be populated under native conditions. Interestingly, the conformational ensembles associated with each cysteine arrangement are both consistent with the available cMII-6 NMR ensemble. While one arrangement stabilizes a central α-helix and lowers the flexibility of the linker, the other arrangement offers an entropical compensation by destabilizing the helix and increasing the flexibility of the linker. This result provides an example of how NcCYP can be used to complement NMR ensembles by generating additional atomistic detail for the native state.

Analysis of the generated conformational ensemble of RTD-1

Figure 2A shows the projections of generated RTD-1 low-energy conformations (no higher than 20 kcal/mol from

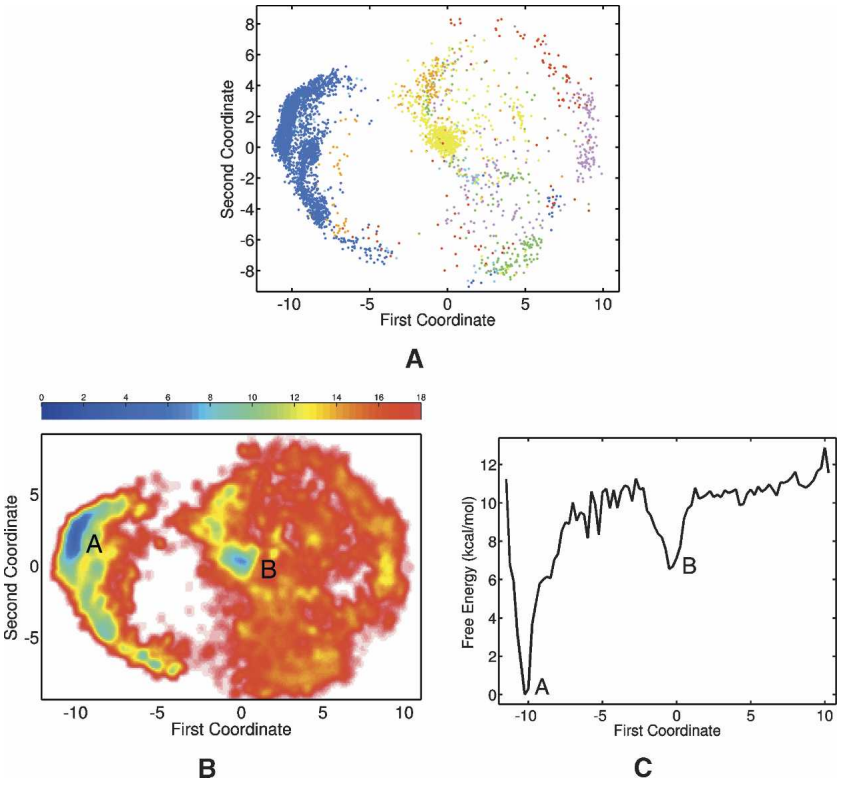

Figure 2. (*A*) 2D landscape obtained with ScIMAP for generated RTD-1 conformations with energies no higher than 20 kcal/mol from the global minimum energy obtained. Each point is color-coded according to the cysteine arrangement in corresponding conformation: blue denotes the 4–17 6–15 8–13 native arrangement observed in the NMR ensemble; sky blue denotes at least one native disulfide bond is present, with the rest of the cysteines unpaired; red denotes 4–17 is formed as under native conditions, with the rest scrambled; yellow denotes the 4–6 8–13 15–17 arrangement; green, plum, and orange denote the remaining all-scrambled arrangements. (*B*) Landscape is color-coded with free energy values (red–blue spectrum indicates high–low values). The lowest free energy state labeled A corresponds to the blue (4–17 6–15 8–13) cluster in *A*. The second-lowest free energy state labeled B corresponds to the yellow (4–6 8–13 15–17) cluster in *A*. (*C*) Plot shows free energy values measured over the first coordinate. The global minimum labeled A is about 10 RT units lower than the local minimum labeled B.

the global minimum energy obtained) onto the first two coordinates obtained with ScIMAP. Each point in this two-dimensional (2D) landscape is color-coded according to the cysteine arrangement in the corresponding conformation as shown in Figure 2A. Cysteine arrangements where not all three disulfide bonds are formed are practically all filtered out when considering only conformations within 20 kcal/mol from the global minimum energy obtained. Including conformations with higher potential energies in this 2D landscape reveals abundant cases where not all three disulfide bonds are formed (data not shown).

Figure 2A reveals two well-separated clusters (in blue and yellow). Generated conformations with cysteines arranged in native disulfide bonds (4–17 6–15 8–13) as in the NMR ensemble are clustered together in the blue cluster. The yellow cluster, albeit smaller, is associated with conformations in the 4–6 8–13 15–17 cysteine arrangement.

Free energy values calculated over this 2D landscape are shown in Figure 2B. A red-to-blue color spectrum denotes high-to-low free energy values. The comparison of Figure 2, A and B, shows that the lowest free energy state (labeled A) corresponds to the blue cluster of projections of conformations with the native 4–17 6–15 8–13 cysteine arrangement. The second-lowest free energy state (labeled B) corresponds to the yellow cluster of projections of conformations with the nonnative 4–6 8–13 15–17 cysteine arrangement.

Figure 2C shows the free energy values as a function of the first ScIMAP-obtained coordinate and allows direct comparison of the free energies of the regions labeled A and B in Figure 2B. Figure 2C shows that the A-labeled region is a global minimum, and the B-labeled region is a local minimum. The free energy difference between these two minima is ~6 kcal/mol (~10 RT units at room temperature). Therefore, the 4–17 6–15 8–13 cysteine arrangement (i.e., the one consistent with the NMR ensemble) is correctly recovered as the native one populated by RTD-1 at equilibrium.

The conformational ensembles corresponding to the two free energy minima are shown in Figure 3. The ensemble corresponding to the global minimum is shown in Figure 3A, whereas that corresponding to the alternative, higher energy, minimum is shown in Figure 3B. Figure 3A shows that the conformations associated with the native state have β-hairpin folds and highly flexible turns connecting well-formed β-sheets. This result confirms the hypothesis by Daly et al. (N.L. Daly, Y.-K. Chen, K.J. Rosengren, U.C. Marx, M.L. Phillips, A.J. Waring, W. Wang, R.I. Lehrer, and D.J. Craik, in prep.) that the turns connecting the sheets are highly flexible. A comparison between these conformations and the NMR ensemble in Figure 1A reveals that the native state

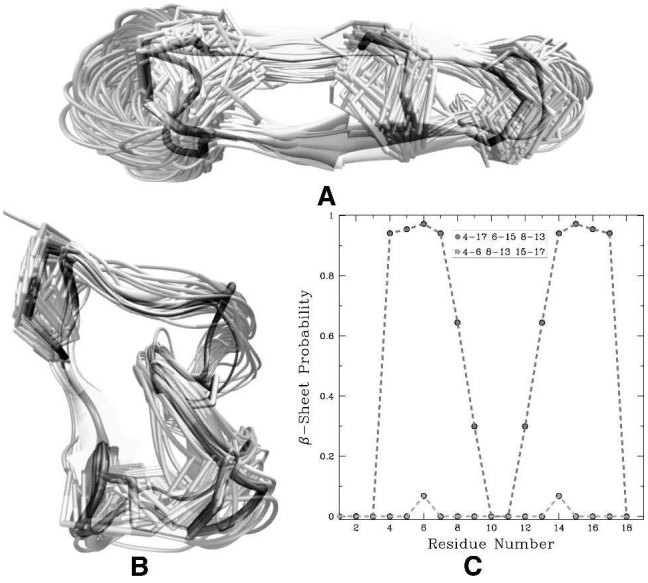

Figure 3. Conformations associated with lowest free energy states labeled A and B (corresponding to free energy values ≤10 kcal/mol) are, respectively, shown in *A* and *B* superimposed in transparent with VMD (Humphrey et al. 1996) over the minimum potential energy conformation among them. (*A*) The β-sheets are well-formed as in the NMR ensemble, with most of the flexibility located in the turns, and cysteines are in a ladder arrangement. (*B*) There is no significant secondary structure among conformations associated with the second-lowest free energy state. (*C*) Dark gray line, showing secondary structure probabilities for each amino acid over ensemble in *A*, reveals well-formed β-sheets. Light gray line, showing probabilities over ensemble in *B*, reveals negligible secondary structure.

predicted by NcCYP, though very similar in fold to the NMR ensemble, has higher structural variability. Figure 3B shows that the nonnative conformations corresponding to the higher-energy minimum are structurally similar to one another, lacking secondary structure, with few of them showing partially formed β-sheets. The structural similarity, considering that these conformations are generated independently of one another by NcCYP, suggests that these conformations are not artifacts of the force field or the method, but rather belong to an actual higher free energy state (that may be temporarily populated during denaturation). This is supported by the fact that conformations representative of the ensembles associated with the two lowest free energy states retain their structural integrity when subjected to equilibration in explicit solvent (for details, see Supplemental Material).

The average secondary structure in the two ensembles (Fig. 3A,B) can be quantified. A probability value is calculated as a Boltzmann average over amino acid secondary structure assignments (obtained with STRIDE) (Frishman and Argos 1995) over each conformation of an ensemble. Figure 3C compares probabilities measured over conformations in Figure 3A (dark gray line) to those measured over conformations in Figure 3B (light gray

line). Figure 3C shows the presence of β-sheets among conformations with the native 4–17 8–13 15–17 cysteine arrangement and the negligible secondary structure among conformations with the nonnative 4–6 8–13 15–17 arrangement. These results show that NcCYP is accurately predicting the exclusive preference of RTD-1 for the 4–17 6–15 8–13 cysteine arrangement and the β-sheet fold under native conditions, fully consistent with the NMR ensemble.

Analysis of the generated conformational ensemble of cMII6

Figure 4A shows the 2D landscape obtained with ScIMAP for the generated cMII-6 conformations with potential energies no higher than 20 kcal/mol from the global minimum energy obtained. Each point in the landscape is color-coded according to the cysteine arrangement in the corresponding conformation as shown in Figure 4A. Cysteine arrangements, where not all two disulfide bonds are formed, are all filtered out when considering only conformations within 20 kcal/mol from the global minimum energy. Cases with unpaired cysteines are observed among conformations with higher energies (data not shown). Figure 4A reveals abundant blue- and green-colored projections and a smaller cluster of yellow-colored ones.

Free energy values calculated as a function of the first two ScIMAP-obtained coordinates are shown in Figure 4B. Comparing Figure 4, A and B, two localized lowest free energy states separated by the $x = 0$ line emerge. These states (labeled A and B in Fig. 4B) appear to correspond to the green- and blue-colored projections in Figure 4A (the respective 8–22 9–14 and 8–14 9–22 cysteine arrangements) and are almost equally probable at room temperature.

Free energy values calculated over the first coordinate, shown in Figure 4C, reveal two minima labeled A and B for direct comparison with Figure 4B. While the B-labeled minimum associated with the 8–14 9–22 arrangement (the arrangement present in the NMR ensemble) has a lower free energy value than the A-labeled minimum

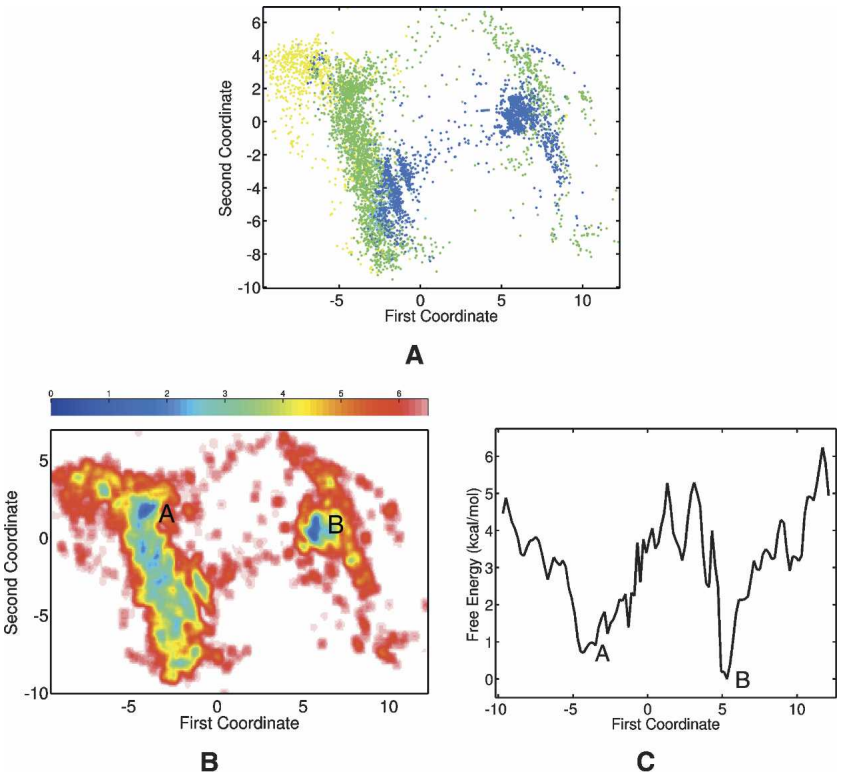

Figure 4. (*A*) 2D landscape obtained with ScIMAP for generated cMII-6 conformations with energies no higher than 20 kcal/mol from the global minimum energy obtained. Each point is color-coded according to the cysteine arrangement in corresponding conformation: blue denotes the 8–14 9–22 native arrangement observed in the NMR ensemble; green and yellow denote the remaining 8–22 9–14 and 8–9 14–22 arrangements, respectively. (*B*) Landscape is color-coded with free energy values (high–low values shown in red–blue color spectrum). The two lowest free energy states A and B seem to, respectively, correspond to green (8–22 9–14) and blue (8–14 9–22) projections. (*C*) Plot shows free energy values measured over the first coordinate. The free energy difference between the two minima is about 1 RT.

associated with the 8–22 9–14 arrangement, the free energy difference between these minima is relatively small (0.7 kcal/mol, i.e., ~1 RT unit at room temperature), indicating that both arrangements can be populated under native conditions. An inspection of the cMII-6 NMR ensemble in Figure 1D shows that, from a purely geometrical consideration, it is not difficult for cMII-6 to accommodate both cysteine pairings.

The separation of the two minima around the $x = 0$ line is structurally meaningful. Inspection of conformations corresponding to the two minima, shown in Figure 5, A and B, reveals that the separation corresponds to the formation of an α-helix. Conformations with the 8–22 9–14 cysteine arrangement, shown in Figure 5A, lack well-formed secondary structure, whereas those associated with the 8–14 9–22 arrangement, shown in Figure 5B, have a central α-helix, similar to the NMR ensemble (Clark et al. 2005) in Figure 1D. This fact is quantified in Figure 5C, which compares secondary structure probabilities calculated over conformations in Figure 5A (dark gray line) to those calculated over conformations in Figure 5B (light gray line). Figure 5C shows the low probability of formation of helical structure in the conformations in Figure 5A. In contrast, Figure 5C shows high helical probabilities for the central amino acids in the conformations shown in Figure 5B.

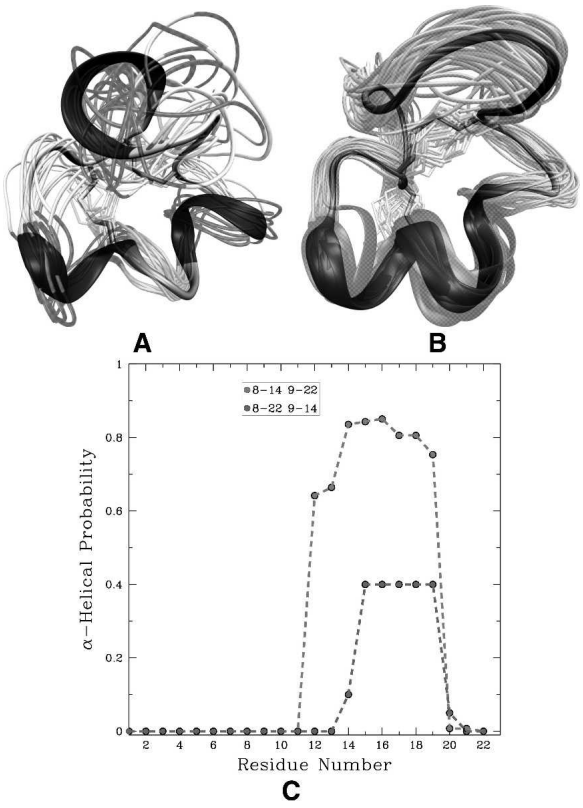

Figure 5. Conformations associated with lowest free energy states labeled A and B (corresponding to free energy values ≤7 kcal/mol) are, respectively, shown in *A* and *B* superimposed in transparent over the minimum potential energy conformation among them using VMD (Humphrey et al. 1996). (*A*) There is no distinguishable helical structure among conformations associated with the free energy state labeled A (8–22 9–14 arrangement). (*B*) A well-formed central α-helix can be seen in conformations associated with the free energy state labeled B (8–14 9–22 arrangement). (*C*) Dark gray line, showing secondary structure probabilities for each amino acid over ensemble in *A*, indicates the lack of secondary structure. Light gray line, showing probabilities over ensemble in *B*, indicates the presence of a well-formed central α-helix in the ensemble.

The structural similarity among conformations in Figure 5A suggests that these conformations are not an artifact of the method. As for the first peptide, this is also confirmed by the fact that, when subjected to explicit solvent equilibrations, conformations representative of the ensembles associated with the two lowest free energy states retain their structural integrity.

The conformational ensembles corresponding to the 8–22 9–14 and 8–14 9–22 cysteine arrangements shown, respectively, in Figure 5, A and B, are both consistent with the cMII-6 NMR ensemble (see Fig. 1D). The 8–14 9–22 arrangement stabilizes the central α-helix also present in the NMR ensemble and so lowers the potential energy of associated conformations. On the other hand, the 8–22 9–14 arrangement, while destabilizing the helix, displays higher linker flexibility, offering an entropic compensation under native conditions. Interestingly, the linker appears highly flexible in the NMR ensemble as well.

The small free energy difference between these two cysteine arrangements leads us to hypothesize that both arrangements can be populated under native conditions. This is not surprising, considering that cMII-6 is engineered from a naturally occurring peptide, hence not necessarily optimized to uniquely fold to a particular structure. Application on this peptide illustrates that, by ranking the feasibility of different native-like conformational substates, NcCYP can direct experimental procedures toward further refinement of the native state ensemble.

Analysis of the generated conformational ensemble of kalata B8

Figure 6A shows the 2D landscape obtained for the generated kalata B8 conformations with potential energies no higher than 20 kcal/mol from the global minimum energy obtained. Each point in the landscape is color-coded according to the cysteine arrangement in the corresponding conformation. Figure 6A reveals that the two highest populated clusters are blue and red: Blue color-codes the native 1–15 5–17 10–23 cysteine arrangement also present in the NMR ensemble; red color-codes arrangements with one native disulfide bond 5–17 and the rest of the cysteines unpaired or scrambled.

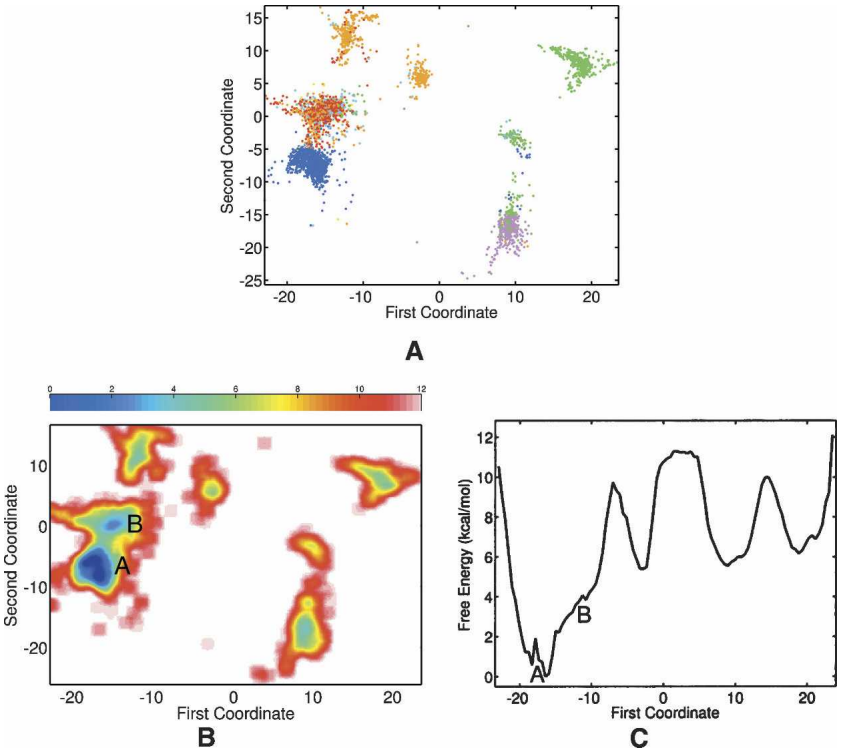

Figure 6. (*A*) 2D landscape obtained with ScIMAP for generated kalata B8 conformations with energies ≤20 kcal/mol from the global minimum energy obtained. Each point is color-coded according to the cysteine arrangement in corresponding conformation: blue denotes the 1–15 5–17 10–23 native arrangement observed in the NMR ensemble; sky blue denotes at least one native disulfide bond is present, with the rest of the cysteines unpaired; yellow denotes 1–15 is formed as under native conditions, with the rest scrambled; red denotes 5–17 is formed as under native conditions, with the rest scrambled; plum and orange denote the remaining all-scrambled arrangements. (*B*) Landscape is color-coded with free energy values (high–low values shown in red–blue color spectrum). Lowest free energy state labeled A corresponds to the blue (1–10 5–17 10–23) cluster in *A*. (*C*) Free energy values measured over the first coordinate show that A is about 4 RT units lower than B.

Free energy values calculated over the landscape are shown in Figure 6B. Comparing Figure 6, A and B, shows that the lowest free energy state (labeled A) corresponds to the blue cluster of projections. The second-lowest free energy state (labeled B) corresponds to the red cluster. The projection of the free-energy landscape on the first coordinate is shown in Figure 6C, where the two minima are labeled accordingly. The free energy difference between the two minima is ~4 kcal/mol (~7 RT units at room temperature). The 1–15 5–17 10–23 cysteine arrangement (the one consistent with the NMR ensemble) is correctly recovered as the native one.

The conformational ensembles corresponding to the two free energy minima A and B are shown in Figure 7, A and B, respectively. Comparing Figure 7 and Figure 1E reveals that the ensemble corresponding to A is very similar to the NMR ensemble but has overall higher structural variability in the loop regions. Interestingly, an α-helix is partially populated in one of the loops in this ensemble. Figure 7B shows that the nonnative conformations corresponding to the higher-energy minimum are also structurally similar to one another, overall lacking secondary structure, with few showing partially formed β-sheets. Conformations representative of the ensembles associated with the two lowest free energy states retain their structural integrity when subjected to equilibration in explicit solvent.

Figure 7C compares the probability of β-sheet formation between the two different minima. Figure 7C shows the presence of β-sheets among conformations with the native 1–15 5–17 10–23 cysteine arrangement and negligible secondary structure among the conformations with the nonnative arrangement. Figure 7D similarly compares the probability of helix formation in each of the ensembles, showing a partial helix formed with low probability in the lowest free energy state. These results suggest that, in agreement with the NMR ensemble, NcCYP is able to capture the exclusive preference of kalata B8 for the 1–15 5–17 10–23 cysteine knot motif, the β-sheet fold, and the high flexibility of the loops connecting the β-strands under native conditions.

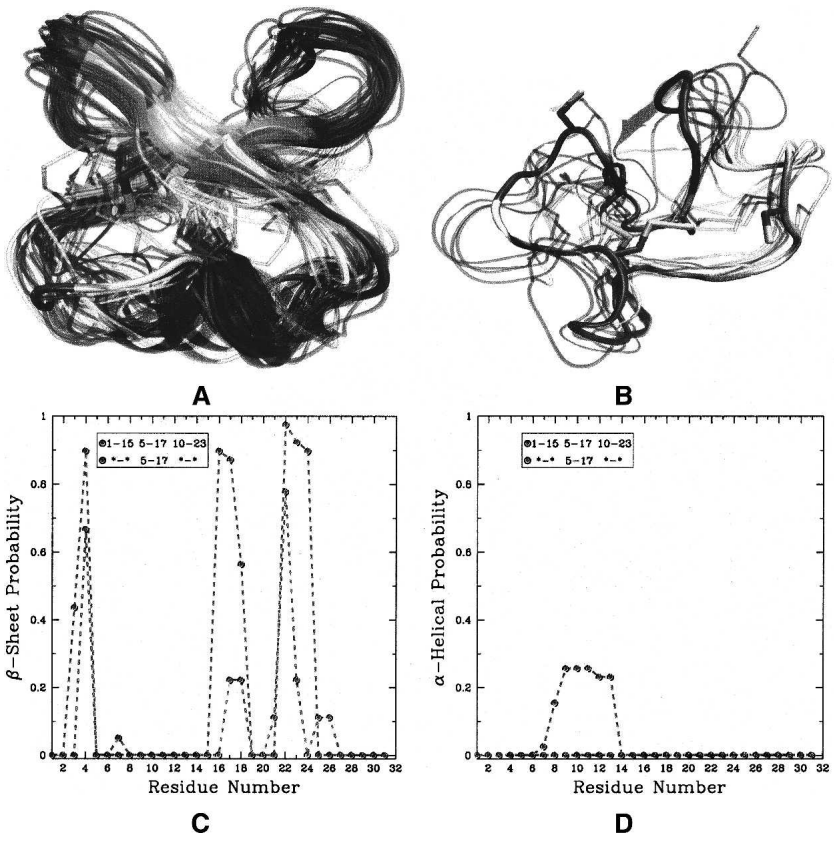

Figure 7. Conformations associated with lowest free energy states labeled A and B (corresponding to free energy values no higher than 8 kcal/mol) are, respectively, shown in *A* and *B* superimposed in transparent with VMD over the minimum potential energy conformation among them. (*A*) β-Sheets are well-formed as in NMR ensemble, with most of the flexibility located in the loop regions. Cysteines are in the knot motif as in NMR ensemble. (*B*) There is no significant secondary structure among conformations associated with the second-lowest free energy state. (*C*) Dark gray line, showing secondary structure probabilities calculated for each amino acid over ensemble in *A*, reveals well-formed β-sheets. Light gray line, showing probabilities obtained over ensemble in *B*, reveals negligible secondary structure. (*D*) Some helicity is observed with low probability in ensemble in *A*.

Application of NcCYP to enhance NMR ensembles

Although beyond the scope of this article, it is worth mentioning that an additional application of NcCYP is the refinement and enhancement of NMR ensembles. As detailed in the Supplemental material, NcCYP can be applied to populate efficiently low-energy structural basins in the energy landscape. In particular, using NMR structures as representative of such basins, NcCYP can generate a large ensemble of new lower-energy conformations. As an example, we have used RTD-1, cMII-6, and kalata B8 NMR structures as starting points to enhance the NMR conformational ensembles of each of these peptides: The native cysteine arrangement is recovered for RTD-1 and kalata B8 NMR structures; for cMII-6, both arrangements predicted by NcCYP are recovered (data not shown). The enhanced NMR ensembles are fully consistent with the results discussed above that were obtained without using structural information.

Discussion

We propose a method (named NcCYP) to predict native conformational diversity of cyclic cysteine-rich peptides using minimal information, namely, amino acid sequence and backbone cyclization. The results show that NcCYP is able to capture in atomistic detail the large conformational ensemble defining the native state and the diversity in native cysteine arrangements.

The conformational ensembles predicted by NcCYP for two naturally occurring peptides, 18 and 31 aa long, are completely consistent with the respective NMR ensembles. The native cysteine arrangement is recovered as the global free energy minimum in each case. Application on an engineered 22-aa sequence reveals two almost equally probable cysteine arrangements. The conformational ensembles associated with each arrangement are consistent with the published NMR ensemble.

NcCYP's multi-resolution hierarchic exploration obtains a detailed view of the large conformational space relevant for the native state. The ability to generate a large number of distinct cyclic conformations, model diversity in cysteine arrangements, and focus the exploration toward increasingly relevant energy minima are key ingredients to the method's success. The main limitation of the method is that it becomes computationally more expensive for longer sequences. In order to overcome this limitation, a coarser resolution than the backbone representation employed in this work could be used at an initial stage to further reduce the dimensionality of the relevant conformational space.

Additionally, approximations associated with the employed force field may affect the results. Different force fields could be used to further refine and test the predicted ensembles. The application presented here shows that it is possible, however, to predict efficiently a few relevant native features from minimal a priori information in reasonable time, at least for a class of cyclic peptides.

Materials and Methods

NcCYP employs a multi-resolution approach to generate cyclic conformations. The method initially uses a low-resolution representation, explicitly modeling only backbone atoms. Employing an inverse kinematics procedure, cyclic coordinate descent (CCD) (Canutescu and Dunbrack 2003), NcCYP efficiently searches the space of backbone angles (ϕ, ψ) and obtains backbone conformations that satisfy the cyclization constraints (for details, see Supplemental material). A large number (hundreds of thousands) of cyclic backbone conformations are obtained independently of one another through a parallel computation framework, as detailed below.

A high-resolution representation is obtained by converting each cyclic backbone to an all-atom conformation. A search for optimal side chains is conducted (Heath et al. 2007) as detailed in the Supplemental material. The resulting all-atom conformation is refined with a physically realistic all-atom energy function. The AMBER9 ff03 force field (Duan et al. 2003) and the implicit Generalized Born (GB) solvation model (Still et al. 1990) are used. A discussion on the choice of the force field is presented in the Supplemental material. In addition to the all-atom refinement, a feasible cysteine arrangement is assigned to each conformation. Proximity and energetic criteria are used to associate an energetically favorable cysteine arrangement to each generated conformation. It is worth stressing that a particular disulfide bond pattern is not enforced a priori. In the end, a large number of low-energy all-atom conformations with distinct cyclic backbone structures and feasible cysteine arrangements are obtained.

This initial ensemble provides a broad view of the conformational space. In the following, we refer to this initial stage of NcCYP as sampling the equilibrium ensemble with dynamic disulfide bond formation (SEEDD). Because of the large number and structural diversity of independently generated conformations, SEEDD can overcome the problem of becoming trapped in false energy minima. Clustering SEEDD-obtained conformations according to their cysteine arrangements is a natural way to reveal populated conformational states associated with the different arrangements.

After this broad view of energy minima, the exploration proceeds iteratively. We refer to this second stage as populate minima (POPMIN). POPMIN uses conformations representative of energy minima as starting points from which to structurally guide the search toward new lower-energy conformations, which are in turn clustered as above to reveal even more minima. This continues until convergence, that is, until no new lower-energy minima appear in successive iterations.

Finally, obtained conformations are subjected to a spatial and energetic analysis. By means of nonlinear dimensionality reduction, the high-dimensional conformational space of generated conformations is reduced to a low-dimensional space spanned by few coordinates. These coordinates reveal conformational clusters and allow us to define a low-dimensional free-energy landscape. This analysis, together with SEEDD and POPMIN, is further discussed below.

SEEDD—a broad view of conformational space

An all-atom cyclic conformation is generated as follows (for more details, see Supplemental material):

(1) An initial conformation for the backbone chain is generated first.
(2) The chain is cyclized by bringing its termini close enough by means of CCD (Canutescu and Dunbrack 2003), as detailed below. A peptide bond is then imposed between the termini.
(3) Energetically feasible side-chain configurations are then added onto the cyclic backbone by following the side-chain reconstruction proposed by Heath et al. (2007).
(4) The resulting all-atom conformation undergoes an energetic refinement and is retained if its potential energy is no higher than 20 kcal/mol of the minimum energy obtained thus far. Otherwise, the search resumes from step 1. The minimum energy is updated if the retained conformation's energy is lower.
(5) Cysteines are arranged in disulfide bonds as described below. If the resulting conformation's potential energy is higher than 20 kcal/mol of the minimum energy obtained at this stage, the search resumes from step 1. Otherwise, the conformation is retained and the minimum energy is updated accordingly.

Imposing cyclization

NcCYP relates the 2n backbone dihedral angles (ϕ, ψ) in a chain of n amino acids through the geometric constraints imposed by cyclization on the N and C termini (step 2 in SEEDD). The cyclization imposes six constraints, three positional and three orientational. Solutions can be enumerated for chains with up to six independent variables (dihedral angles) (Wedemeyer and Scheraga 1999). Early work on cyclic molecules first sampled arbitrary values for the independent variables 1 through $2n - 6$, then solved exact algebraic equations for the six remaining variables $2n - 5$ to $2n$ (Go and Scheraga 1970). Instead of solving exactly for short subchains with at most six variables, later methods such as Random Tweak (Fine et al. 1986) and CCD (Canutescu and Dunbrack 2003) framed the closure of arbitrarily long chains as an optimization problem. Closing long chains remains an active area of research (Kolodny et al. 2005). Our previous work employed CCD to explore extensively the $2n - 6$ dimensional conformational space of protein fragments

of length n (Shehu et al. 2006). In this work, NcCYP uses CCD to explore the conformational space of cyclic peptides 18–31 aa long.

Formation of disulfide bonds

NcCYP does not enforce a specific disulfide bond pairing between cysteines. Both proximity and energetic criteria are used to find a feasible cysteine arrangement for each conformation (step 5 in SEEDD). First, cysteines closer than a certain threshold are identified. Disulfide bonds are then formed between cysteine pairs and optimized through a short energy minimization. Disulfide bonds are allowed to break and form iteratively until convergence (see Supplemental material). It is worth noting that even if specific cysteine pairings were enforced a priori, this would not significantly reduce the dimensionality of the conformational space. On the other hand, considering all possible disulfide bond patterns is not practical. The number of ways to pair all given 2k cysteines in k disulfide bonds is $\frac{\prod_{i=0}^{k-1}\binom{2k-2i}{2}}{k!}$ (e.g., 15 pairings for six cysteines). Requiring that all cysteines be paired imposes a priori information on the actual native state. On the other hand, enumerating all possible cysteine arrangements, allowing for unpaired cysteines, is also not practical, as the number of arrangements is $1 + \sum_{j=1}^{k} \frac{\prod_{i=0}^{j-1}\binom{2k-2i}{2}}{j!}$ (e.g., 76 arrangements for six cysteines). For these reasons, the method used in this work avoids explicit enumeration (see Supplemental material). The same force field and solvation model are used to determine the feasibility of a cysteine arrangement; no particular terms in the AMBER ff03 force field promote formation of disulfide bonds.

A parallel computation framework

The complete independence in the generation of one conformation from another makes the computation intrinsically parallel. Parallelization allows us to sample a very large number of conformations (the entire computation for the results presented in this article takes about a week when distributed among 50 CPUs).

POPMIN—an iterative exploration of the native basin

SEEDD-obtained conformations are clustered according to cysteine arrangements to reveal those arrangements associated with energy minima. Few (1 2) lowest-energy (i.e., with energies less than 5 kcal/mol from the minimum) conformations are selected to represent an energy minimum associated with a particular cysteine arrangement. These conformations, deemed seeds, are used as reference structures from which lower-energy conformations can be generated (for details, see Supplemental material). When a large number (1000–3000) of conformations are obtained starting from a particular cysteine arrangement, the obtained conformations are clustered as above to reveal potentially lower-energy minima from which to start another iteration. If a newly generated conformation has an energy value lower than 1.0 kcal/mol from the seed used to generate it, and it is further than 1.0 Å lRMSD from the seed, then the conformation is considered as the new representative, and it replaces the seed. When seeds do not change between successive iterations, i.e., no lower-energy states emerge, POPMIN is considered converged.

Spatial analysis of generated conformations

The high-dimensional conformational space populated by NcCYP-generated conformations is projected onto a low-dimensional space by means of ScIMAP, a nonlinear dimensionality reduction method (Das et al. 2006; Plaku et al. 2007). ScIMAP computes the nearest-neighbors graph by connecting each conformation to its nearest neighbors. The shortest distance between two conformations is defined as the length of the shortest path that connects them in the nearest-neighbor graph. The shortest-path distances between L conformations selected as landmarks and the remaining conformations are computed and stored in a matrix M. The top eigenvectors of the matrix M are used as an orthogonal base set for the low-dimensional projection of conformations. We apply ScIMAP as in the method of Das et al. (2006) to obtain a few coordinates that span the space of generated conformations. Different numbers of landmarks (1000–2000) and nearest neighbors (20–30) have been tested to ensure accuracy and robustness of the obtained projections.

Free energy analysis of generated conformations

Free energy values are calculated on the low-dimensional conformational landscape by using a modified version of WHAM (Ferrenberg and Swendsen 1988, 1989). The modification takes into account that the NcCYP-generated conformations are not obtained with a constant temperature constraint; therefore, they do not define a canonical ensemble. Assuming that the conformational space is sampled uniformly (Shehu et al. 2007b) and that the sampling is dense, we divide the low-dimensional space in cells and associate a density to each cell. The potential energy associated with conformations whose projections fall on a particular cell on the landscape is averaged to smooth out the noise in the force field or solvation model used. Different grid cell sizes have been tested to ensure the robustness of the results.

Acknowledgments

We thank Erion Plaku and Hernan Stamati for ScIMAP-related assistance; Ilya Yildirim, Carlos Simmerling, and David Case for AMBER9-related assistance; and John Stone for visualization of cyclic proteins with VMD. This work is supported by the National Science Foundation (C.C., career grant no. CHE-0349303; and L.E.K. and C.C., grant no. CCF-0523908), National Institutes of Health (L.E.K., grant no. GM078988), Welch Foundation (C.C., Norman Hackermann Young Investigator award and grant no. C-1570), and Sloan Foundation (L.E.K.). Equipment used was funded by NSF grant no. CNS-0421109 and grant no. CNS-0454333 in partnership between Rice University, AMD and Cray. A.S. is partly supported by a fellowship from the Nanobiology Training Program of the W.M. Keck Center for Computational and Structural Biology of the Gulf Coast Consortia (National Institutes of Health grant no. 1 R90 DK71504-01).

References

Abkevich, V.I. and Shakhnovich, E.I. 2000. What can disulfide bonds tell us about protein energetics, function, and folding: Simulations and bioinformatics analysis. *J. Mol. Biol.* **300:** 975–985.

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181:** 223–230.

Baneyx, F. and Mujacic, M. 2004. Recombinant protein folding and misfolding in *Escherichia coli*. *Nat. Biotechnol.* **22:** 1399–1408.

Canutescu, A.A. and Dunbrack Jr., R.L. 2003. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* **12:** 963–972.

Clark, R.J., Fischer, H., Dempster, L., Daly, N., Rosengren, K.J., Nevin, S.T., Meunier, F.A., Adams, D.J., and Craik, D.J. 2005. Engineering stable peptide toxins by means of backbone cyclization: Stability of the α-conotoxin MII. *Proc. Natl. Acad. Sci.* **102:** 13767–13772.

Craik, D.J. 2006. Seamless proteins tie up their loose ends. *Science* **311:** 1563–1564.

Craik, D.J., Cemazar, M., and Daly, N.L. 2006. The cyclotides and related macrocyclic peptides as scaffolds in drug design. *Curr. Opin. Drug Discov. Devel.* **9:** 251–260.

Czaplewski, C., Stanislaw, O., Liwo, A., and Scheraga, H.A. 2004. Prediction of the structures of proteins with the UNRES force field, including dynamic formation and breaking of disulfide bonds. *Protein Eng. Des. Sel.* **17:** 29–36.

Daly, N.L., Clark, R.J., Plan, M.R., and Craik, D.J. 2006. Kalata B8, a novel antiviral circular protein, exhibits conformational flexibility in the cystine knot motif. *Biochem. J.* **393:** 619–626.

Das, P., Moll, M., Stamati, H., Kavraki, L.E., and Clementi, C. 2006. Low-dimensional free energy landscapes of protein folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci.* **103:** 9885–9890.

Dombkowski, A.A. and Crippen, G.M. 2000. Disulfide recognition in an optimized threading potential. *Protein Eng. Des. Sel.* **13:** 679–689.

Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G.M., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., et al. 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24:** 1999–2012.

Eisenmesser, E.Z., Millet, O., Labeikovsky, W., Korzhnev, D.M., Wolf-Watz, M., Bosco, D.A., Skalicky, J.J., Kay, L.E., and Kern, D. 2005. Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438:** 117–121.

Ferré, F. and Clote, P. 2005. DiANNA: A web server for disulfide connectivity prediction. *Nucleic Acids Res.* **33:** W230–W232. doi: 10.1093/nar/gki412.

Ferrenberg, A.M. and Swendsen, R.H. 1988. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.* **61:** 2635–2638.

Ferrenberg, A.M. and Swendsen, R.H. 1989. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **63:** 1185–1198.

Fine, R.M., Shenkin, P.S., Wang, H.J., Yarmush, D.L., and Levinthal, C. 1986. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. *Proteins* **1:** 342–362.

Fiser, A., Cserzo, M., Tudos, E., and Simon, I. 1992. Different sequence environments of cysteines and half cystines in proteins. Application to predict disulfide forming residues. *FEBS Lett.* **302:** 117–120.

Frishman, D. and Argos, P. 1995. Knowledge-based protein secondary structure assignment. *Proteins* **23:** 566–579.

Go, N. and Scheraga, H.A. 1970. Ring closure and local conformational deformations of chain molecules. *Macromolecules* **3:** 178–187.

Heath, A.P., Kavraki, L.E., and Clementi, C. 2007. From coarse-grain to all-atom: Towards multiscale analysis of protein landscapes. *Proteins* **68:** 646–661.

Hilser, V.J., Oas, T., Dowdy, D., and Freire, E. 1998. The structural distribution of cooperative interactions in proteins: Analysis of the native state ensemble. *Proc. Natl. Acad. Sci.* **95:** 9903–9908.

Hogg, P.J. 2003. Disulfide bonds as switches for protein function. *Trends Biochem. Sci.* **28:** 210–214.

Humphrey, W., Dalke, A., and Schulten, K. 1996. VMD—visual molecular dynamics. *J. Mol. Graph.* **14:** 33–38. http://www.ks.uiuc.edu/Research/vmd/.

Karplus, M. and Kuriyan, J. 2005. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci.* **102:** 6679–6685.

Kolodny, R., Guibas, L., Levitt, M., and Koehl, P. 2005. Inverse kinematics in biology: The protein loop closure problem. *Int. J. Robot. Res.* **24:** 151–163.

Krause, K., Pineda, L.F., Peteranderl, R., and Reissman, S. 2000. Conformational properties of a cyclic peptide bradykinin B-2 receptor antagonist using experimental and theoretical methods. *J. Pept. Res.* **55:** 63–71.

Lindorff-Larsen, K., Best, R.B., DePristo, M.A., Dobson, C.M., and Vendruscolo, M. 2005. Simultaneous determination of protein structure and dynamics. *Nature* **433:** 128–132.

Loiseau, N., Gomis, J.M., Santolini, J., Delaforge, M., and Andre, F. 2003. Predicting the conformational states of cyclic tetrapeptides. *Biopolymers* **69:** 363–385.

Martelli, P.L., Fariselli, P., and Casadio, R. 2004. Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. *Proteomics* **4:** 1665–1671.

Mucchielli-Giorgi, M.H., Hazout, S., and Tuffery, P. 2002. Predicting the disulfide bonding state of cysteines using protein descriptors. *Proteins* **46:** 243–249.

Onuchic, J.N., Luthey-Schulten, Z., and Wolynes, P.G. 1997. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **48:** 545–600.

Palmer, A.G.I., Kroenke, C.D., and Loria, J.P. 2001. Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods Enzymol.* **339:** 204–238.

Plaku, E., Stamati, H., Clementi, C., and Kavraki, L.E. 2007. Fast and reliable analysis of molecular motions using proximity relations and dimensionality reduction. *Proteins* **67:** 897–907.

Quik, M., Polonskaya, Y., Kulak, J., and McIntosh, J.M. 2001. Vulnerability of ^{125}I-α-conotoxin MII sites to nigrostriatal damage in monkey. *J. Neurosci.* **21:** 5494–5500.

Rayan, A., Senderowitz, H., and Goldblum, A. 2004. Exploring the conformational space of cyclic peptides by a stochastic search method. *J. Mol. Graph. Model.* **22:** 319–333.

Schnell, J.R., Dyson, H.J., and Wright, P.E. 2004. Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu. Rev. Biophys. Biomol. Struct.* **33:** 119–140.

Shehu, A., Clementi, C., and Kavraki, L.E. 2006. Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. *Proteins* **65:** 164–179.

Shehu, A., Kavraki, L.E., and Clementi, C. 2007a. On the characterization of protein native state ensembles. *Biophys. J.* **92:** 1503–1511.

Shehu, A., Clementi, C., and Kavraki, L.E. 2007b. Sampling conformation space to model equilibrium fluctuations in proteins. *Algorithmica* **48:** 303–327.

Shin, S.Y., Yoo, B., Todaro, L.J., and Kirshenbaum, K. 2007. Cyclic peptoids. *J. Am. Chem. Soc.* **129:** 3218–3225.

Skolnick, J., Kolinski, A., and Ortiz, A.R. 1997. MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265:** 217–241.

Still, W.C., Tempczyk, A., Hawley, R.C., and Hendrickson, T. 1990. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112:** 6127–6129.

Tang, Y.-Q., Yuan, J., George, O., Osapay, K., Tran, D., Miller, C.J., Ouellette, A.J., and Selsted, M.E. 1999. A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated θ-defensins. *Science* **286:** 498–502.

Trabi, M., Schirra, H.J., and Craik, D.J. 2001. Three-dimensional structure of RTD-1, a cyclic antimicrobial defensin from Rhesus macaque leukocytes. *Biochemistry* **10:** 4211–4221.

Wedemeyer, W.J. and Scheraga, H.J. 1999. Exact analytical loop closure in proteins using polynomial equations. *J. Comput. Chem.* **20:** 819–844.