John von Neumann Institute for Computing

The Role of Flexible Stem Geometries in Protein Loop Structure Prediction

M. Mönnigmann, C. A. Floudas

published in

NIC Workshop 2006,
From Computational Biophysics to Systems Biology,
Jan Meinke, Olav Zimmermann,
Sandipan Mohanty, Ulrich H.E. Hansmann (Editors)
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **34**, ISBN-10: 3-9810843-0-6,
ISBN-13: 978-3-9810843-0-6, pp. 115-118, 2006.

http://www.fz-juelich.de/nic-series/volume34

The Role of Flexible Stem Geometries in Protein Loop Structure Prediction

M. Mönnigmann[1] **and C. A. Floudas**[2]

[1] Process Systems Engineering
RWTH Aachen University, Templergraben 55, 52056 Aachen
E-mail: moennigmann@lpt.rwth-aachen.de

[2] Department of Chemical Engineering,
Princeton University, Princeton, NJ 08544-5263
E-mail: floudas@titan.princeton.edu

Previous works on loop structure prediction treat the loop reconstruction problem, that is, the geometry of the protein into which the loop must fit is assumed to be given. In ab initio protein structure prediction, however, this information is not available, but loop structure and the anchoring stem region structure have to be predicted simultaneously. The resulting loop structure prediction problem with flexible stems treated here is considerably more difficult than the reconstruction problem.

The proposed approach is based on (i) dihedral angle sampling, (ii) structure optimization by energy minimization with a physically based energy function, and (iii) clustering. In contrast to previous works, clustering is not only used to identify conformers that are likely to be close to the native structure, but also to identify far-from-native decoys.

The method is tested on a large test set. Surprisingly, the average of the smallest rmsd found in ensembles of 2000 conformers depends linearly on the length of the sequence. Since our test set is large, different methods for selecting close-to-native conformers from ensembles can be compared in a meaningful way.

1 Introduction

Without the great structural flexibility of loops, many proteins could not fold into compact structures. Since loops posses great geometric flexibility, loop structure is difficult to predict. Recent approaches assume information on the location of the anchoring residues is available, that is, the loop reconstruction problem is treated[1]. While using information on the geometry of the anchoring residues may be justified in homology modeling, approaches of this type are ruled out in *ab initio* prediction. In order to benchmark the prediction precision that can be achieved in *ab initio* loop structure prediction, we apply a methodology that predicts both, the geometry of the anchoring residues, and the loop structure itself, to a large test set.

2 Methods

Conformers are generated with probability functions in a discretized (ϕ, ψ)-space similar to those used by DePristo et al.[2]. We distinguish between three types of residues, (i) α-helical amino acids as defined by DSSP code H, (ii) β-strand (DSSP code E), and (iii) loop[3]. To qualify as a loop an amino acid sequence must not be of DSSP type E or H, not be at either terminal, and must be located between strands or helices. Probability distributions

n_r [-]	10	12	14	16	18	20
rmsd [Å]	.9842	1.412	1.825	2.159	2.451	2.772

Table 1. Average of smallest rmsd in ensembles of given loop length. The symbol n_r denotes the number of residues in a loop including the stem residues. Rmsds depend linearly on the loop length[1]. Results for odd loop lengths are omitted for brevity.

are determined by investigating a reference set of known structures. This set comprises all proteins with experimental resolution of 2.2Åor better in the PdbSelect25 set[3].

In order to take the influence of the anchor regions into account, we add three stem residues at both ends and subject these additional six residues to the same prediction procedure as the loop residues. We optimize side chain angles using the Dunbrack rotamer library[4]. Energies are calculated with the ECEPP/3 force field[5]. After the side chain optimization the entire structure is subjected to an energy minimization with NPSOL[6].

Clustering methods have been used before to identify groups of conformers that are *likely* to be similar to each other and, ultimately, *likely* to be similar the native structure[1]. In contrast, our clustering approach identifies conformers that are likely *not* to be similar to the unknown native structure, and discards those structures. As a result the overall quality of the ensemble is improved, and any strategy for picking out the best conformer in the remaining ensemble is more likely to identify a conformer that is close to native. The clustering method used here is based on the observation that cluster size and rmsd to the native structure are correlated[1].

3 Results

We applied our method to 3215 loops extracted from the PdbSelect25 set of proteins and to a set of 65 loops from the CASP6 targets[1]. Due to space restrictions, we can only report a fraction of our results[1].

Table 1 assesses the quality of the conformers generated by dihedral angle sampling. These results are obtained by determining the minimum rmsd to native for all conformers generated by dihedral angle sampling, and subsequently averaging over the minimum rmsds of all ensembles of loops of a given length. From the data in table 1 we infer that dihedral angle sampling does not restrict the prediction accuracy[1].

Figure 1 summarizes some of our prediction results[1]. Results labeled *average rmsds by energy* are determined by identifying the lowest energy conformer in each ensemble, recording its rmsd to native, and subsequently averaging over the rmsds for all ensembles of loops of a given length. The average rmsds by colony energy are determined correspondingly, where the candidate conformer in each ensemble is picked by identifying the lowest colony energy[7]. The remaining data shown in Figure 1 is obtained by applying our clustering algorithm[1].

From Figure 1 it is evident that the conformer that generates the largest cluster after any clustering step $k = 0, 1, 2$ is on average a better prediction for the native structure than the lowest energy or lowest colony energy conformer. Cluster size is therefore a better criterion for the identification of good conformers than energy or colony energy. Moreover, the largest cluster conformer found after step $k + 1$ of the clustering algorithm is as good

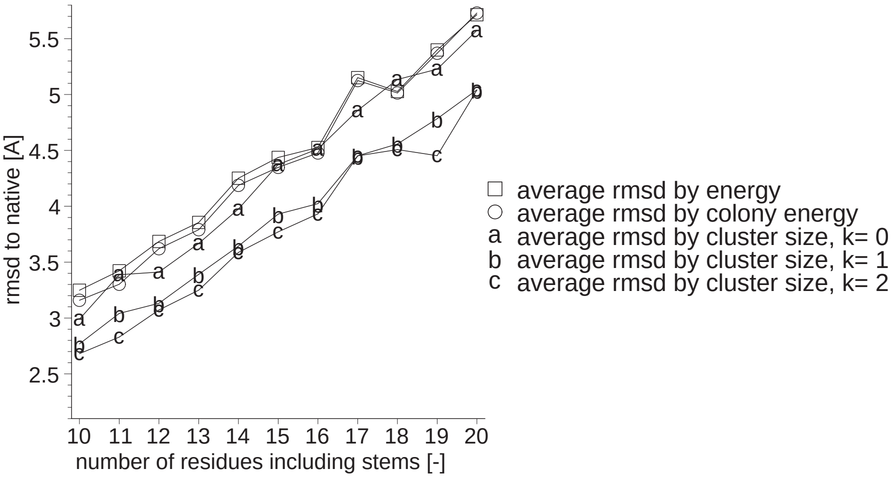

Figure 1. Average of rmsds as a function of total loop length for the loops from the PdbSelect25 proteins. Lines are added as a guide to the eye.

as, or better than, the largest cluster conformer found in step k. In this sense, repeatedly applying the clustering algorithm improves the largest cluster conformer on average.

For the large set of loops we investigated, the rmsd predicted by cluster size after iterative clustering, averaged over the ensembles of all loops of the same length, is a linear function of loop length. This indicates that the prediction quality is currently not limited by the dihedral angle sampling, but by the strategy for selecting a conformer that is likely to be close to native[1]. This holds also for the results for the CASP6 targets[1].

4 Summary and Outlook

We briefly summarized results obtained with a new methodology that allows structure prediction of loops with flexible stem residues. The proposed methodology was applied to a large test set, allowing meaningful results of methods to select conformers from ensembles that are close to the native structure. We compared several selection criteria, namely ECEPP/3[5] energy, colony energy[7], and cluster size before and after application of a new clustering algorithm[1]. The comparison shows that that energy is approximately as good as colony energy, cluster size before applying our clustering approach is on average better than both energy and colony energy, and cluster size after applying our clustering approach is the best criterion.

The loop prediction method developed here is ultimately going to be used in an existing ab initio protein prediction approach[8].In this context, the loop prediction method must not assume information on the surrounding protein to be given, but loops and the remaining parts of the structure must be predicted simultaneously.

Acknowledgments

CAF gratefully acknowledges funding from the National Institute of Health (R01-GM-52032, R24-GM-069736). MM gratefully acknowledges funding by a Deutsche Forschungsgemeinschaft research fellowship (MO 1086/1-1).

References

1. M. Mönnigmann and C. A. Floudas. Loop structure prediction with flexible stem geometries. *Proteins: Structure, Function, and Bioinformatics*, 61:748–762, 2005.
2. M. A. DePristo, P. I. W. de Bakker, S. C. Lovell, and T. L. Blundell. Ab initio construction of polypeptide fragments: Efficient generation of accurate, representative ensembles. *Proteins: Structure, Function, and Bioinformatics*, 51:41–55, 2003.
3. W. Kabsch and Ch. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.
4. M. J. Bower, F. E. Cohen, and R. L. Dunbrack. Prediction of protein side chain rotamers from a backbone dependent rotamer library: a new homology modeling tool. *Journal of Molecular Biology*, 267:1268–1282, 1997.
5. G. Némethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *Journal of Physical Chemistry*, 96:6472–6484, 1992.
6. P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright. *NPSOL 4.0 User's Guide*. Systems Optimization Laboratory, Dept. of Operations Research, Stanford University, CA., 1986.
7. Z. Xiang, C. Soto, and B. Honig. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7432–7437, 2002.
8. J. L. Klepeis and C. A. Floudas. ASTRO-FOLD: A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical Journal*, 85:2119–2146, 2003c.