110a Protein Loop Structure Prediction with Flexible Stem Geometries

Martin Monnigmann and Christodoulos A. Floudas

The importance of loops for the overall three-dimensional structure and function of proteins has been pointed out before (see, e.g., [1]). Despite their short length, loops are of major importance. Without loops, many proteins could not fold into compact structures, and loops are often exposed to the surface and contribute to active and binding sites. Since loops possess greater structural flexibility than strands and helices, and since they have relatively few contacts with the remainder of the structure, loop structure is considerably more difficult to predict than the geometrically regular ß-strands and helices.

Loop structure prediction can be considered a mini ab initio protein folding problem. Some of the methods that have been established for protein structure prediction [2] are ruled out for loop structure prediction, however. Comparative modeling is not suited for loop structure prediction, because the correlation between amino acid sequence and three-dimensional structure is only weak for short sequences [3]. In fold recognition methods, loops often contribute to the overall rmsds in a disproportionate way and limit the prediction quality [4,5]. The need for higher precision loop structure predictions in the context of comparative modeling has motivated recent research on first principles based methods for loops [4-7].

Recent progress in loop structure prediction has been achieved with approaches that combine dihedral angle sampling, steric clash detection, clustering, and scoring or energy function evaluation to build up ensembles of loop conformations, and to select representative structures from those ensembles. Methods for loop structure prediction that are based on dihedral angle sampling and building up ensembles are not new [8]. Much of the recent progress can be attributed to the availability of greater computer power that, among other things, allows for finer discretizations of the dihedral angle space, as well as for sampling more conformers.

All recent papers on loop structure prediction have considered the loop reconstruction problem. In this problem, the anchor geometry of the protein into which the loop must fit is assumed to be known. In contrast, this work addresses the structure prediction of loops with flexible stem residues. While the secondary structure of the stem residues is assumed to be known, the geometry of the protein into which the loop must fit is considered to be unknown in our methodology. As a consequence, the compatibility of the loop with the remainder of the protein is not used as a criterion to reject loop decoys. The loop structure prediction with flexible stems must be considered more difficult than fitting loops into a known protein structure in that a larger conformational space has to be covered. The main focus of this work is to assess the precision of loop structure prediction that can be attained if no information on the protein geometry is available.

The proposed approach is based on (i) dihedral angle sampling, (ii) structure optimization by energy minimization with a physically based energy function, (iii) clustering, and (iv) a comparison of strategies for the selection of loops identified in (iii). Steps (i) and (ii) have similarities to previous approaches to loop structure prediction with fixed stems. Step (iii) is based on a new iterative approach to clustering that is tailored for the loop structure prediction problem with flexible stems. In this new approach, clustering is not only used to identify conformers that are likely to be close to the native structure, but clustering is also employed to identify far-from-native decoys. By discarding these decoys iteratively, the overall quality of the ensemble and the loop structure prediction is improved. Step (iv) provides a comparative study of criteria for loop selection based on energy, colony energy, cluster density, and a hybrid criterion introduced here. The proposed method is tested on a large set of 3215 loops from proteins in the PdbSelect25 set and to 179 loops from proteins from the Casp6 experiment.

We assess the quality of the ensembles generated by dihedral angle sampling and energy optimization. This assessment is done separately from the evaluation of different methods for selecting low rmsd conformers. To our surprise we found that the average of the smallest rmsd found in ensembles of 2000 conformers depends linearly on the length of the sequence. This observation holds for all loop lengths treated, that is, for loops of length 10 (4 loop and 6 stem residues) through length 20 (14 loop and 6 stem residues). Since we gathered statistics for a large set of 3394 loops extracted from the PdbSelect25 set of proteins and the Casp6 targets, this result clearly is not an artifact of selecting prediction targets that give favorable results. This result is surprising, since the conformational space to be sampled grows exponentially.

We compare several methods for selecting conformers from the ensemble that are close to the native structure, namely potential energy, colony energy [9], and cluster size before and after application of the new clustering algorithm. The comparison shows that that energy is approximately as good as colony energy, cluster size before clustering is on average better than both energy and colony energy, and cluster size after iterative clustering is the best criterion. Furthermore we find that the clustering algorithm proposed here does not discard those conformers that have the lowest energies or colony energies. This confirms that conformers that are ranked high with established methods are not discarded by our new clustering approach.

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. The slope of this linear function is, however, larger than the slope of the average smallest rmsd as a function of loop length. Furthermore, there is a gap between these two linear functions. 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. The slope of the average smallest rmsd as a function of loop length is recovered if we allow for five conformers to be selected rather than only one from each ensemble. A hybrid strategy that selects the lowest energy, the lowest colony energy conformer, and the centroids of the largest clusters after k= 0, 1, 2 clustering steps gives the best results and reduces the gap to the average smallest rmsds.

The loop prediction method developed here is ultimately going to be used in the context of an existing ab initio protein prediction [10,11]. 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.

[1] Fiser, A., Do, R. K. G., and Sali, A., Modeling of loops in protein structures. Protein Science 9:1753--1773, 2000.

[2] Floudas, C. A., Fung, H. K., McAllister, S. R., Mönnigmann, M. and Rajgaria, R., Advances in Protein Structure Prediction and De Novo Protein Design: A Review. Chem. Eng. Sc., in print. [3] Cohen, B. I., Presnell, S. R., and Cohen, F. E., Origins of structural diversity within sequentially identical hexapeptides. Protein Science 2:2134--2145, 1993.

[4] Jacobson, M. P., Pincus, D. L., Rappa, C. S., Day, T. J. F., Honig, B., Shaw, D. E., and Friesner, R. A., A hierarchical approach to all-atom protein loop prediction. Proteins: Structure, Function, and Bioinformatics 55:351--367, 2004.

[5] Li, X., Jacobson, M. P., and Friesner, R. A. High-resolution prediction of protein helix positions and orientations. Proteins: Structure, Function, and Bioinformatics 55:368--382, 2004.

[6] DePristo, M. A., de Bakker, P. I. W., Lovell, S. C., and Blundell, T. L. Ab initio construction of polypeptide fragments: Efficient generation of accurate, representative ensembles. Proteins: Structure, Function, and Bioinformatics 51:41--55, 2003.

[7] Rohl, C. A., Strauss, C. E. M., Chivian, D., and Baker, D., Modeling structurally variable regions in homologous proteins with Rosetta. Proteins: Structure, Function, and Bioinformatics 55:656--677, 2004.

[8] Bruccoleri, R. E. and Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137--168, 1987.

[9] Xiang, Z., Soto, C., and Honig, B. 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.

[10] M. Mönnigmann and Floudas, C. A., Protein loop structure prediction with flexible stem geometries. Submitted to Proteins, 2005.

[11] Klepeis, J. L. and Floudas, C. A., 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, 2003.