Exact Enumeration of Protein Conformations from Fragment Assembly

Julian Lee
Department of Bioinformatics and Life Science, Soongsil University, Seoul, Korea
E-mail: jul@ssu.ac.kr

Abstract. Efficient sampling of protein conformations is very important in computational studies of protein folding. The search space of the protein conformations can be reduced by utilizing bioinformatics. The fragment assembly is one such method, where the local structures are selected from the fragment sets prepared using bioinformatics tool. The conformational space of proteins is reduced to a finite discrete set by using fragment assembly. The search space can be further reduced by introducing a consistency condition at the junction of the fragments. An algorithm for exact enumeration of such conformations are introduced, which can be used for generating candidates of the native structure for a relatively short protein.

1. Introduction
Prediction of the native structure of a protein is an important and challenging problem in computational biophysics. According to Anfinsen's thermodynamic hypothesis[1], the native structure is the conformation minimizing the free energy under physiological condition. However, computation of protein native structure by purely physical means has not been so successful, not only because it is difficult to construct a free energy function with reasonable accuracy, but also because the conformational space that has to be sampled is huge. The fragment assembly method, which has exhibited excellent performance in protein structure prediction for sequences with little overall sequence similarity with those in the structural database, is a method that uses bioinformatics method to complement the physico-chemical calculations[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. In a fragment assembly method of protein structure prediction, candidates of possible local structure are constructed for each part of the protein, by collecting fragments from structural database using similarity of sequence features. The conformations are generated by assembling these fragments and only the global tertiary packing of fragments is determined by energy optimization, and hence the name fragment assembly method. By using fragments, not only the burden of modeling interactions between neighboring residues is eliminated, but also the conformational space is reduced to a finite set of discrete conformations, making the conformational search much more tractable.

Despite remarkable successes of fragment assembly methods[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13], the conformational space of a protein has been mostly sampled using stochastic methods coupled with appropriate score function to select the model structure. In this work, we develop an algorithm for exact enumeration of all possible conformations generated by fragment assembly, after achieving a further reduction of the conformational space by implementing additional
constraints for fragment assembly. The exact enumeration algorithm is then tested for a small protein 1m2zB.

The exact enumeration of protein conformations, even for short proteins, is useful in several aspects. Since all the conformations are examined for a given protein, the exact global minimum of a free energy function can be found. For a training set consisting of proteins of known structures, this information can be used to develop an accurate free energy function so that a native-like conformation is produced as the global minimum. Also, method for collecting candidate fragments from structural database may be developed, so that there are many native-like conformations in the search space. Lastly, benchmark tests on various stochastic sampling methods can be performed.

2. Methods
2.1. Structural Database
A list of chains with mutual sequence identities less than 25%, X-ray resolution lower than 3 Å and R factor lower than 1.0 was taken from the protein sequence culling server PISCES [14]. The chains with unknown structure regions were removed to obtain the final database consisting of 1,927 protein chains with a total of 378,149 residues. The fragments are prepared from this database.

2.2. Fragment Set
For each residue of a protein sequence, we collect a set of $k$ fragments from the structural database, corresponding to the most probable conformations of the local neighborhood. By tuning the parameter $k$, one can control the size of the conformational space. The criterion for the fragment selection is the similarity of sequence feature. However, instead of directly comparing the raw sequences, the profile obtained from PSI-BLAST [15] search is compared.

When PSI-BLAST [15] search is performed against a protein sequence database, the sequences similar to the query is retrieved and aligned, and the rate of substitution of each residue of the query protein to other amino acids is then calculated, which is called the sequence profile. The PSI-BLAST sequence profile is a matrix of size (protein length) $\times$ 20, which contains evolutionary information that cannot be obtained from the raw sequence only.

For each residue of a query sequence, a window of size 15 centered on this residue is constructed. The part of the sequence profile lying within this window is considered as the sequence feature vector. Thus, the feature vector is the matrix of size 15 $\times$ 20, except for windows centered on terminal residues where the windows have to be truncated [16, 17, 18, 19, 20].

Similar construction is performed for each window of each protein sequence in the reference dataset. Then $k$ such windows whose feature vectors are closest to the query window is collected. Since $k$ windows are collected from the reference dataset with known structures, the corresponding fragments of size 15 form the fragment set for the query window. The distance measure used for the sequence feature comparison is

$$ d_{ij} = \sum_{i=1}^{15} \sum_{j=1}^{20} |P_{ij} - P'_{ij}| $$

where $P_{ij}$ and $P'_{ij}$ are the elements of the feature vectors [16, 17, 18, 19, 20], except for small details such as fragment sizes or similarity measures used for the sequence feature comparison.

2.3. Fragment Assembly
We assemble fragments to construct a protein conformation. Since a protein chain of length $N$ is assembled from $N$ fragments, and there are $k$ candidate fragments centered on each residue, the total number of possible conformations is $k^N$ if the fragments are assembled without any
restriction. In the current work, the fragments are joined only when they overlap and share at least one residue with similar values of dihedral angles. Two sets of dihedral angles \((\phi_1, \psi_1)\) and \((\phi_2, \psi_2)\) are considered to be similar to each other if either

\[ |\phi_1 - \phi_2| + |\psi_1 - \psi_2| \leq 2^\circ. \tag{2} \]

If we find such a residue, then the second fragment is joined smoothly to the first one starting from this residue. This additional constraint prevents the appearance of any local structure around fragment junction that is not present in the structural database, and reduces the conformational space substantially, while still allowing for generating native-like structures. The fragments that can be joined are precomputed, and then all possible conformations are enumerated.

### 3. Results

To test the exact enumeration algorithm, we performed exact enumeration of the conformations of 1m2z chain B, with sequence length 21 consisting of mostly of \(\alpha\) helices, a protein included in the structural database. For fairness of the test, it is removed from the structural database before the fragment selection, so that the remaining homology with the structural database is sequence identity less than 25 % by construction.

The total number of conformations are given in Table I for various values of \(k\). It is a non-decreasing function of \(k\), as it should be, since the conformational space for a given value of \(k\) always includes the space for smaller value of \(k\) as a subset.

**Table 1.** The total number of fragment-assembled protein conformations, the number of conformations with RMSD < 4.0 Å, their ratio, the smallest value of RMSD found among the conformations, and CPU time, as functions of number of candidate fragments \(k\), for the protein 1m2z chain B.

| \(k\) | total number | < 4.0 | fraction | RMSD(smallest) | CPU time (sec) |
|------|--------------|-------|----------|----------------|----------------|
| 1    | 0            | 0     | –        | –              | 0.06           |
| 2    | 0            | 0     | –        | –              | 0.06           |
| 3    | 381          | 41    | 0.11     | 3.89           | 0.11           |
| 4    | 1360         | 66    | 0.049    | 3.83           | 0.24           |
| 5    | 3232         | 196   | 0.061    | 3.70           | 0.46           |
| 6    | 50414        | 4083  | 0.081    | 3.57           | 6.34           |
| 7    | 164232       | 13950 | 0.085    | 3.57           | 20             |
| 8    | 418175       | 36225 | 0.086    | 2.72           | 52             |
| 9    | 668856       | 55811 | 0.086    | 2.72           | 84             |
| 10   | 1082077      | 94664 | 0.086    | 2.72           | 136            |

As the number of conformations increases, the number of native-like conformations will also increase. It is crucial for the success of a fragment-based protein structure prediction method, that a sufficient number of native-like conformations are generated by fragment assembly. The native-like property of a protein conformation is usually measured by root-mean-square-deviation(RMSD) of its backbone atom coordinates from those of the native structure. The smallest RMSD value found among all the conformations, and the number of conformations with RMSD < 4.0 Å, are displayed in Table I. Note that for \(k = 10\), the total number of conformation would be \(10^{21}\) if the fragments were assembled without imposing any constraint. In the current algorithm, the number of conformations is reduced to about \(10^6\), achieving a
10^{15} fold reduction, but conformations with RMSD as low as 2.72 Å are generated. The CPU time spent for computation in Intel i7 64 bit Hexacore was 136 seconds for generating these conformations.

4. Discussions
In this work, we developed an algorithm for generating all possible conformations assembled from structural fragments. The crucial elements are the additional reduction of the conformational space by imposing constraints at the fragment junctions, and the precomputation of fragment that can be joined together. The algorithm was tested for a small protein 1m2zB. For \( k = 10 \) the number of conformations are reduced by \( 10^{15} \) fold, but conformations with RMSD as small as 2.72 Å were generated. By combining the current algorithm with accurate score function to select the model structure, a novel protein structure prediction algorithm may be developed.

4.1. Acknowledgments
The author thanks Seung-Yeon Kim for useful discussions. This work was supported by the Soongsil University Research Fund.

References
[1] Anfinsen C B 1973 Science 181 223
[2] Jones D T 2005 Proteins 61 (S7) 143
[3] Rohl C, Strauss C, Misura K and Baker D 2004 Methods Enzymol. 383 66
[4] Chikenji G, Fujitsuka Y and Takada S 2006 J. Proc. Acad. Natl. Sci. U.S.A. 103 3141
[5] Lee J, Kim S-Y, Joo K., Kim I and Lee J 2004 Proteins 56 704
[6] Lee J, Kim S-Y and Lee J 2005 Biophys. Chem. 115 209
[7] Lee J, Kim S-Y and Lee J 2005 J. Korean Phys. Soc. 46 707
[8] Kim S-Y, Lee W and Lee J 2006 J. of Chem. Phys. 125 194908
[9] Kim T K and Lee J 2008 J. Korean Phys. Soc. 52 137
[10] Lee D-S, Seok C and Lee J 2008 J. Korean Phys. Soc. 52 1137
[11] Lee J, Lee D-S, Park H, Coutsias E A and Seok C 2010 Proteins 78 3428
[12] Ko J, Lee D-S, Park H, Coutsias E A, Lee J and Seok C 2011 Nucleic Acid Res 39 W210
[13] Park H, Ko J, Joo K., Lee J, Seok C and Lee J 2011 Proteins 79 2725
[14] Wang G and Dunbrack R L 2003 Bioinformatics 19 1589
[15] Altschul S F, Madden T L, Schaffer A A, Zhang J., Zhang Z, Miller W and Lipman D J 1997 Nucleic Acid Res 25 3389
[16] Kim S-Y, Sim J and Lee J 2006 Lecture Notes in Bioinformatics 4115 562
[17] Sim J, Kim S-Y and Lee J 2005 Bioinformatics 21 2844
[18] Joo K, Lee J, Kim S-Y , Kim I, Lee S J and Lee J 2004 J. Korean Phys. Soc. 44 599
[19] Joo K, Kim I, Lee J, Kim S-Y, Lee S J, and Lee J 2004 J. Korean Phys. Soc. 45 1441
[20] Lee J 2009 J. Korean Phys. Soc. 54 1