Description of the KDD-Cup 2004 Protein Data

Author: Ron Elber, Department of Computer Science, Cornell University

The data are generated by the program LOOPP (Learning Observing and Outputting Protein Patterns) that is a product of the group of Ron Elber, Computer Science, Cornell University. The LOOPP server is accessible via
http://ser-loopp.tc.cornell.edu/cbsu/loopp.htm

LOOPP is a fold recognition program. It accepts as input a protein sequence and computes numerous protein features based on the input sequence (for detailed description of the features see below). The different features are combined to assign a global score of similarity to other fully annotated proteins, suggesting likely templates for modeling. LOOPP is also building atomically detailed models using the protein templates with the highest score of (global) similarity to the probe sequence.

Roughly, the features are divided as follow. We consider general properties such as sequence similarity, sequence-to-structure matching, secondary structure fitness, exposed surface area, and matching to the whole sequence family of the probe sequence (so called profile) to the template. Each of these properties is examined in five different ways. We compute the raw sequence, reverse energy, and Z score, and special threading energy and Z scores according to the alignment of the current feature.

The program above is trained against numerous sets in which we try (known) probe sequences against templates, learning to identify similar folds.

One of the training set of LOOPP includes 304 proteins and was discussed in detail in the manuscript: Octavian Teodorescu, Tamara Galor, Jaroslaw Pillardy, and Ron Elber, "Enriching the sequence substitution matrix by structural information", Proteins, Structure, Function and Genetics, 54:41-48(2004). For the purpose of KDD cup the set was divided into two. One part was used for training and the other part was used for testing.

The input provided to the KDD cup was the score of individual features. Below I provide a detailed explanation of all the features that were released to the KDD Cup. They are given in the same order they were released. However the running index is as in the original test file for easier comparison (available at http://www.tc.cornell.edu/research/cbsu/webresources/tmp/blastuser/var/loopp_old.htm) The KDD set was the Octavian third test with low shuffle of the Z score.

Features

9. Length of alignment. We align locally the probe sequence against the template. The longer is the alignment (given in percentage of the length of the probe sequence) the better is the match

10. Percentage of sequence identity. Percentage of aminoacids that are identical after alignment. The higher, the better the match.

12. Z score for global sequence alignment (using the BLOSUM matrix with structurally dependent gaps)

13. Reverse energy for global sequence alignment (i.e. align the probe sequence against the template and the reverse of the probe sequence against the template – a model of a random sequence and subtract the two.

14. The raw score of global sequence alignment (opposite sign to the usual score, i.e. negative is better)

15. A raw score of the TE 13 energy function that tests fitness of sequence to structure. The energy is computed according to the global sequence alignment of feature 14. (reference -- Dror Tobi and Ron Elber, "Distance dependent, pair potential for protein folding: Results from linear optimization", Proteins, Structure Function and Genetics, 41, 40-16 (2000)).

16. The Z score of the TE13 energy function. The rest the same as 14.

Features 12-16 include our standard five measures of similarity for a given feature. The Z score, the reverse energy, the raw score, TE13 energy, and TE13 Z score. This set of five is used in other measures as well. The Z score is the most sensitive and accurate, however, it is the most expensive to compute. In some applications we therefore avoid computing it. Most of the properties discussed are using both local and global alignments.

17-21. Scores of local sequence alignment with the same measures of similarity as described in features 12-16.

22-26 Scores of global threading alignment (threading is matching a sequence into a template structure) using the five similarity measures mentioned above (reference - Jaroslaw Meller and Ron Elber, "Linear Optimization and a double Statistical Filter for protein threading protocols", Proteins, Structure, Function and Genetics, 45,241-261(2001)

27-31 Scores of Local threading feature (with the same five measures of similarity)

32-36 Scores from global alignments using a mixed substitution matrix that includes both sequence and structure similarity measures (reference - Octavian Teodorescu, Tamara Galor, Jaroslaw Pillardy, and Ron Elber, "Enriching the sequence substitution matrix by structural information", Proteins, Structure, Function and Genetics, 54:41-48(2004)).

37-41. The same mixed substitution matrix as in 32-36 but this time the scores are computed with local alignment.

42-46. Scores from three way mixed feature (a substitution matrix using sequence similarity, structure matching, and secondary structure prediction). Global alignment. Five measures as above.

47-51 Three way mixed feature as in 42-46. This time scores are computed with local alignment. Five measures as above.

52-56 Scores of global alignment of predicted secondary structure of the probe sequence against the known secondary structure of the template. Five measures as before.

57-61 Scores of local alignment of predicted secondary structure of the probe sequence. Five measures as before.

62-66 We generate a profile for family of the probe sequence and aligned it (globally) against the template (a profile is the probability of having an amino acid at a particular position using multiple sequence alignment of the probe sequence family). Scores of the five measures (as before) are provided.

67-71. Scores of profile alignments as in 62-66, this time with local alignment.

72-76. Score of alignment of solvent exposed surface area (predicted from probe sequence versus the known template), five measures of similarity, global alignment.

76-81 Same as 72-76. This time scores of local alignments of exposed surface areas.

82. A measure of secondary structure fitness based on the position of the secondary structure elements

83. Another measure of secondary structure score.