Protein Fold Recognition Using an Optimal Structure-Discriminative Amino Acid Index

J. Ben Rosen[1]. Robert H. Leary[2]. Per Jambeck[3]. Connie X. Wu[4].

Keywords: protein folding, fold recognition, combinatorial extension, supervised learning, amino acid index

1 Introduction.

Identifying the fold class of a protein sequence of unknown structure is a fundamental problem in modern biology. We have applied a supervised learning algorithm FoldID [1] to the classification of protein sequences of length 64 to 300 with low sequence identity from a library of 174 structural classes created by the Combinatorial Extension (CE) structural alignment methodology [2,3]. A class of rules is considered that assigns test sequences to structural classes based on the closest match of an amino acid index profile of the test sequence, which is a numerical vector of dimension N for sequences of length N, to a numerical profile centroid vector for each class. A mathematical optimization procedure is applied to determine an amino acid index of maximal structural discriminatory power by maximizing the ratio of between-class to within-class profile variation. The optimal index is computed as the solution to a generalized eigenvalue problem, and its performance for fold classification of aligned sequences in the training library is compared to that of other published indices using cross-validation techniques.

The algorithm is also tested on raw, unaligned sequences using a combination of local and global alignment techniques to align the numerical profile vectors corresponding to raw sequences with the various class centroid vectors.

2 Computational Results.

For aligned sequences in the CE training library, the optimal index has significantly more structural discriminatory power than all currently known indices in the AAindex database [4], including average surrounding hydrophobicity, which it most closely resembles (Figure 1). It demonstrates more than 70% cross validation classification accuracy on the structurally aligned sequences in the CE fold library over all folds, and nearly 100% accuracy on many individual folds with distinctive conserved structural features such as the EF hand fold family of calcium binding proteins.

For unaligned raw sequences, the corresponding numerical profile vector is first aligned with each class centroid using both local and global dynamic programming techniques. Sufficiently high scoring subsequences that match particular fold class centroid profiles are tentatively assigned to that structural class. The method has proved to be successful in many cases for properly classifying

[1] Dept. of Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0114, USA. E-mail: jbrosen@cs.ucsd.edu

[2] San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0505, USA. E-mail: leary@sdsc.edu

[3] Dept. of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093, USA. E-mail: jambeck@bioeng.ucsd.edu

[4] San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0505, USA. E-mail: cxwu@sdsc.edu

subsequences of the raw sequences on which the CE library is based and extracting the CE structural alignments based only on the sequence, particularly where the number of gaps and/or deletions is relatively small. Tests on raw sequences outside the CE library are encouraging, with successful structural identifications (as determined by subsequent CE analysis) having been made for structural classes with highly conserved features. For example, all seven test instances of EF hand structures not in the original CE library were correctly identified and aligned using FoldID.

3 Figure.

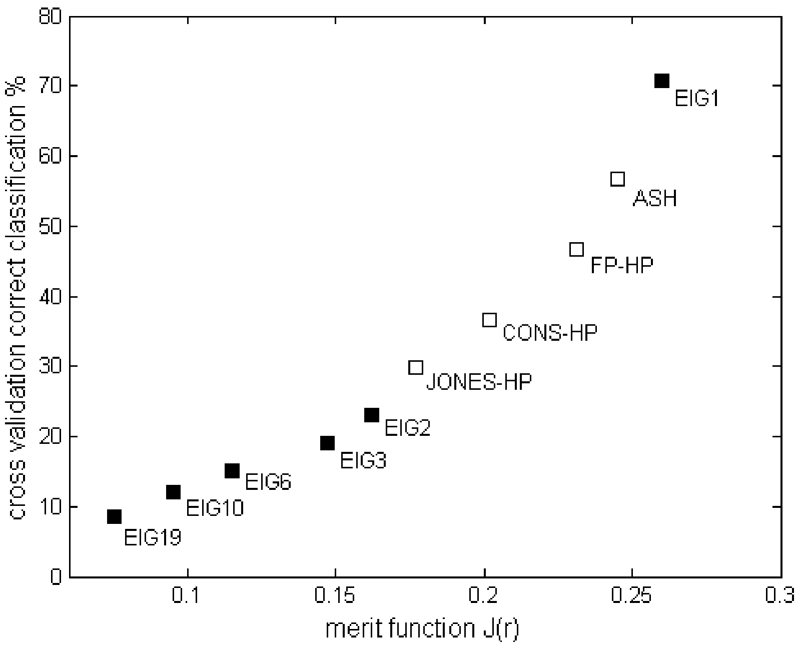

Figure 1. Relative performance of various amino acid indices for CE fold classification. Each EIGi index is computed by FoldID as the i-th eigenvector of the generalized eigenproblem. The remaining indices are various members of the hydrophobicity family.

4 References and bibliography.

[4] Kawashima, S. and Kanehisa, M. 2000. AAindex: amino acid index database. *Nucleic Acids Research* 28:334.

[1] Leary, R.H., Rosen, J.B. and Jambeck, P. 2004. An optimal structure-discriminative amino acid index for protein fold recognition. *Biophysical Journal* 86:411-419.

[2] Shindyalov, I.N. and Bourne, P. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* 11:739-747.

[3] Shindyalov, I.N. and Bourne, P. 2000. An alternative view of protein fold space. *Proteins* 38:247-260.