AAindex : A Database of Amino Acid Indices and Mutation Matrices

K. Tomii
tomii@kuicr.kyoto-u.ac.jp

M. Kanehisa
kanehisa@kuicr.kyoto-u.ac.jp

Institute for Chemical Research, Kyoto University
Uji, Kyoto 611 Japan

Abstract

An amino acid index, which is a set of 20 numerical values, represents the various properties of amino acids. A similarity matrix, also called a mutation matrix, which is a set of 20 × 20 numerical values, represents the similarity between amino acids, and is used for protein sequence alignments and similarity searches. We have collected 402 amino acid indices and 42 published mutation matrices, and organized the database named AAindex which is made publicly available on the Internet.

1 Introduction

The amino acid sequence analysis often provides important insights into the tertiary structure and biological function of proteins. Sequence alignments are the basic strategy to infer three-dimensional structural similarity and/or functional similarity. The amino acid similarity matrix, also called a mutation matrix, is the basis of protein sequence alignments.

As reported previously (Nakai *et al.* [1]) we constructed and maintain a database of amino acid indices. The database, AAindex, now contains 402 published indices with the result of single-linkage hierarchical cluster analysis. Here we make a new addition to the database: a collection of 42 amino acid mutation matrices.

Starting from the efforts of Dayhoff *et al.* [2] who were the first to compile such a mutation matrix, there have been reports of various matrices for use in search of protein sequence similarity and structure prediction. Dayhoff's matrix is based on the observation of accepted point mutation frequency in closely related sequences. Recently attempts were made to directly observe amino acid exchanges from more divergent sequences, and also to compute for different structural classes from structurally known proteins. Mutation matrices were also constructed from amino acid indices.

富井健太郎，金久實：京都大学化学研究所，〒 611 京都府宇治市五ケ庄

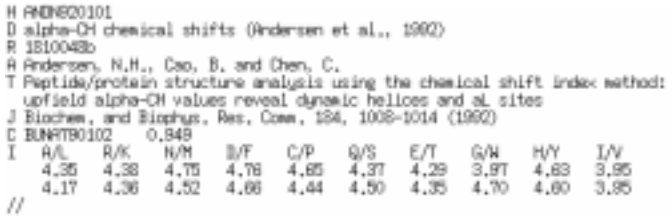

Figure 1: An example of the entry of the AAindex database

2 Format of the Database

AAindex is organized in a flat-file format with one entry corresponding to one index(matrix), i.e., a set of 20(210) numerical values, and associated reference information. A sample entry of the database is shown in Figure 1. Each record of an entry is identified by the following codes: H, accession number; D, data description; R, LITDB identifier; A, author(s); T, title of article; J, journal reference; C, accession numbers of similar entries with the correlation coefficients of 0.8(-0.8) or more(less); I, actual data in the specified order; and *, optional comments.

3 Accessing the Database

The database is publicly made available by the Japanese GenomeNet database service at the following address:

ftp://ftp.genome.ad.jp/db/genomenet/aaindex

When the service is accessed through Gopher and WWW, a database entry may be obtained by using the DBGET Integrated Database Retrieval System.

Acknowledgement

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Informatics', from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] K. Nakai, A. Kidera, M. Kanehisa, "Cluster Analysis of Amino Acid Indices for Prediction of Protein Structure and Function," *Protein Eng.*, Vol. 2, pp. 93-100, 1988.

[2] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, "A Model for Evolutionary Change in Proteins," *Atlas of Protein Sequence and Structure.*, (M.O. Dayhoff ed.) Vol. 5, suppl. 3, pp. 345-358, 1978.