368–369 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

AAindex: Amino Acid Index Database

Shuichi Kawashima, Hiroyuki Ogata and Minoru Kanehisa*

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Received September 8, 1998; Accepted October 15, 1998

ABSTRACT

AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids. It consists of two sections: AAindex1 for the amino acid index of 20 numerical values and AAindex2 for the amino acid mutation matrix of 210 numerical values. Each entry of either AAindex1 or AAindex2 consists of the definition, the reference information, a list of related entries in terms of the correlation coefficient, and the actual data. The database may be accessed through the DBGET/LinkDB system at GenomeNet (http://www.genome.ad.jp/dbget/) or may be downloaded by anonymous FTP (ftp://ftp.genome.ad.jp/db/genomenet/aaindex/).

INTRODUCTION

The variety and specificity of protein three-dimensional structures and biological functions are due to the combination of the 20 different amino acids as specified by the genetic code. The amino acids are the building blocks of proteins each having different characteristics in terms of the shape, the volume, and the chemical reactivity among others. A large body of experimental and theoretical research has been performed to characterize physicochemical and biochemical properties of individual amino acids. The derived property is often represented by a set of 20 numerical values that is called the amino acid index.

In addition to the properties of individual amino acids, the relations between amino acids are also represented by numerical values in the analysis of protein sequences and structures. Especially, the amino acid mutation matrix, also called the amino acid similarity matrix, is the basis for optimization in protein sequence alignments and similarity searches. The amino acid mutation matrix is generally a set of 20×20 numerical values, or a set of 210 numerical values since the matrix is usually symmetric. The AAindex database is a collection of published amino acid indices and mutation matrices.

BACKGROUND

In 1988 Nakai *et al*. collected 222 amino acid indices from research papers and investigated the relationships by the hierarchical cluster analysis (1). They identified four major classes, α-helix and turn propensities, β-strand propensity, hydrophobicity that can further be divided into subclasses, and other physicochemical properties such as bulkiness of amino acid residues. In 1996 Tomii and Kanehisa (2) increased the size of the collection to include 402 indices and re-performed the clustering. The result was generally in good agreement with the previous work, but for the sake of convenience the collection was divided into six major classes: α and turn propensities, β propensity, amino acid composition, hydrophobicity, physicochemical properties, and other properties.

Tomii and Kanehisa (2) also collected 42 amino acid mutation matrices from the literature and conducted extensive analysis on the correlations among them and with the amino acid indices. The AAindex database was initiated by Nakai *et al*. (1), was expanded by Tomii and Kanehisa (2), and is continuously updated by the present authors.

THE CURRENT DATABASE

The AAindex database is a flat file database that consists of two sections: AAindex1 for the amino acid indices and AAindex2 for the amino acid mutation matrices. The format of the two sections is as follows.

AAindex1

The AAindex1 section currently contains 434 amino acid indices. A sample entry of AAindex1 is shown in Figure 1. Each entry consists of an accession number, a short description on the index, the reference information, and the numerical values for the property of 20 amino acids. In addition, it contains neighbor information; namely, the cross-links to other entries with an absolute value for the correlation coefficient of 0.8 or larger. With the links the user can identify a set of entries describing similar properties. In some instances the values are not reported for all 20 amino acids. When available we adopt the estimates by Kidera *et al*. (4) who tried to fill missing values by statistical considerations. When the estimates were not available, the missing values were either replaced by the mean value of the rest or simply filled with zeros.

AAindex2

The AAindex2 section currently contains 66 amino acid mutation matrices: 47 symmetric matrices and 19 non-symmetric matrices. A sample entry of AAindex2 is shown in Figure 2. The format of the entry is almost the same as that of AAindex1 except that it contains 210 numerical values (20 diagonal and $20 \times 19/2$ off-diagonal elements) for a symmetric matrix and 400 or more

*To whom correspondence should be addressed. Tel: +81 774 38 3270; Fax: +81 774 38 3269; Email: kanehisa@kuicr.kyoto-u.ac.jp

```
H PTIO830101
D Helix-coil equilibrium constant (Ptitsyn-Finkelstein, 1983)
R 0904057
A Ptitsyn, O.B. and Finkelstein, A.V.
T Theory of protein secondary structure and algorithm of its prediction
J Biopolymers 22, 15-25 (1983)
* Charged state for Arg, His, Lys, Asp, and Glu
C ROBB760103    0.903  QIAN880109    0.886  QIAN880108    0.884
  SUEM840101    0.877  QIAN880111    0.857  QIAN880107    0.846
  QIAN880110    0.835  FAUJ880102    0.832  FINA770101    0.826
  ROBB760104    0.817  QIAN880131   -0.826  ISOY800104   -0.832
  QIAN880132   -0.833  CHOP780213   -0.835  GEIM800108   -0.840
  CHAM830101   -0.841  LEVM780106   -0.854  CHOP780216   -0.855
  PRAM900104   -0.858  LEVM780103   -0.860  QIAN880133   -0.864
  GEIM800111   -0.876  QIAN880135   -0.899  QIAN880134   -0.920
I   A/L    R/K    N/M    D/F    C/P    Q/S    E/T    G/W    H/Y    I/V
    1.10   0.95   0.80   0.65   0.95   1.00   1.00   0.60   0.85   1.10
    1.25   1.00   1.15   1.10   0.10   0.75   0.75   1.10   1.10   0.95
//
```

Figure 1. An example of the amino acid index entry in the AAindex database (AAindex1). Each record of an entry is identified by the one-letter codes: H, accession number; D, definition of the entry; R, LITDB (3) literature database identifier; A, author(s); T, title of the journal article; J, journal citation information; 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.

numerical values for a non-symmetric matrix (some matrices include a gap or distinguish two states of cysteine).

AVAILABILITY

The AAindex database can be retrieved through the DBGET/LinkDB system (5) of the Japanese GenomeNet service (6) at http://www.genome.ad.jp/dbget/

The DBGET/LinkDB system integrates most of the major molecular biology databases and is especially suited for using hyperlinks to related entries within the AAindex database as well as to the other databases.

Alternatively, the entire database may be copied and used locally. The URL for anonymous FTP is: ftp://ftp.genome.ad.jp/db/genomenet/aaindex/

Users are requested to cite this article when making use of the AAindex database.

ACKNOWLEDGEMENTS

We thank Drs Kenta Nakai and Kentaro Tomii for the initial developments of the AAindex database. This work was supported in part by the Grant-in-Aid for Scientific Research on the Priority Area ‘Genome Science’ from the Ministry of Education, Science, Sports and Culture of Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

```
H NIEK910101
D Structure-derived correlation matrix 1 (Niefind-Schomburg, 1991)
R 1713140
A Niefind, K. and Schomburg, D.
T Amino acid similarity coefficients for protein modeling and sequence
  alignment derived from main-chain folding angles
J J. Mol. Biol. 219, 481-497 (1991)
C NIEK910102    0.998  TUDE900101    0.809
I Data ordered by I+J*(J-1)/2 where I,J = ARNDCQEGHILKMFPSTWYV
    1.00   0.09   1.00  -0.29  -0.10   1.00   0.05  -0.08   0.30   1.00
   -0.35  -0.17   0.04  -0.10   1.00   0.32   0.05  -0.14   0.11  -0.16
    1.00   0.70   0.09  -0.24   0.07  -0.34   0.39   1.00  -0.40  -0.18
    0.37   0.08   0.12  -0.28  -0.45   1.00  -0.18   0.15   0.04  -0.01
   -0.04   0.08  -0.14   0.06   1.00  -0.25   0.05  -0.06  -0.28  -0.15
   -0.13  -0.14  -0.23   0.01   1.00   0.51   0.00  -0.23   0.01  -0.27
    0.29   0.57  -0.54  -0.09   0.25   1.00   0.55   0.14  -0.19   0.13
   -0.25   0.16   0.50  -0.34  -0.23  -0.17   0.31   1.00   0.39  -0.10
   -0.13  -0.09  -0.25   0.22   0.36  -0.29  -0.08   0.15   0.41   0.05
    1.00  -0.18   0.15  -0.13  -0.31  -0.05  -0.25  -0.24  -0.05   0.17
    0.27  -0.08  -0.16  -0.02   1.00   0.10  -0.18  -0.10   0.02   0.13
   -0.18  -0.08   0.10  -0.32  -0.53  -0.25   0.04  -0.31  -0.28   1.00
   -0.38  -0.10   0.04   0.02   0.31  -0.23  -0.41   0.28  -0.05  -0.34
   -0.61  -0.18  -0.48  -0.13   0.35   1.00  -0.59  -0.12   0.01  -0.18
    0.12  -0.21  -0.50   0.10   0.10   0.30  -0.31  -0.33  -0.31   0.19
   -0.21   0.33   1.00  -0.11   0.00  -0.21  -0.19   0.00  -0.17  -0.12
   -0.10  -0.01   0.03  -0.12   0.02  -0.19   0.15  -0.03   0.05   0.07
    1.00  -0.59  -0.01   0.04  -0.19   0.25  -0.28  -0.57   0.19   0.13
    0.13  -0.46  -0.43  -0.31   0.24  -0.13   0.34   0.46   0.06   1.00
   -0.23  -0.03  -0.15  -0.22  -0.10  -0.03  -0.09  -0.24  -0.11   0.67
    0.23  -0.13   0.08   0.13  -0.42  -0.25   0.35  -0.04   0.15   1.00
//
```

Figure 2. An example of the amino acid mutation matrix entry in the AAindex database (AAindex2). The data format is the same as described in Figure 1. The order of the matrix elements may be computed by the equation or examined in the database documentation file.

REFERENCES

1 Nakai,K., Kidera,A. and Kanehisa,M. (1988) *Protein Engng.*, **2**, 93–100.
2 Tomii,K. and Kanehisa,M. (1996) *Protein Engng.*, **9**, 27–36.
3 Seto,Y., Ihara,S., Kohtsuki,S., Ooi,T. and Sakakibara,S. (1988) In Lesk,A.M. (ed.), *Computational Molecular Biology*. Oxford University Press, New York, pp. 27–37.
4 Kidera,A., Konishi,Y., Oka,M., Ooi,T. and Scheraga,H.A. (1985) *J. Protein Chem.*, **4**, 23–55.
5 Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y. and Kanehisa,M. (1998) *Pacific Symp. Biocomput., 1998*, 683–694.
6 Kanehisa,M. (1997) *Trends Biochem. Sci.*, **22**, 442–444.