BIOINFORMATICS APPLICATIONS NOTE Vol. 14 no. 4 1998
Pages 376–377

Self-organizing tree growing network for classifying amino acids

Huai-chun Wang[1,2], Joaquin Dopazo[3] and Jose Maria Carazo[1]

[1]Centro Nacional de Biotecnologia-CSIC, Universidad Autonoma, 28049 Madrid, Spain, [2]Institute of Medical Information (AMMS), 27 Taiping Road, 100850 Beijing, China and [3]Glaxo Wellcome, C/ Severo Ochoa 2, 28760 Tres Cantos, Madrid, Spain

Received on October 30, 1997; revised on December 15, 1997; accepted on December 18, 1997

Abstract

Summary: *A self-organizing tree growing neural network was applied to classify amino acids and amino acid exchange matrices.*

Availability: *SOTA, is freely available by anonymous FTP or at http://www.cnb.uam.es/~bioinfo/Software/sota.*

Contact: *carazo@cnb.uam.es*

Most protein sequence analysis tasks rely on a measure of similarity between different amino acids. There are at least 13 published scoring matrices for amino acids, based on genetic codes, physicochemical properties, observed frequency of mutations, secondary structural matching and structural properties (Johnson and Overington, 1993). Various attempts have been made to group amino acids based on these matrices, including dendrograms (Sneath, 1966; Doolittle, 1979; Johnson and Overington, 1993; Jones *et al.*, 1994), Venn diagrams (Dickerson and Geis, 1969; Taylor, 1986; Taylor and Jones, 1993), principal components analysis (Johnson and Overington, 1993), multidimensional projection (Jones *et al.*, 1992, 1994; Taylor and Jones, 1993) and Sammon's non-linear mapping (Agrafiotis, 1997). In particular, Johnson and Overington (1993) used both hierarchical clustering by the KITSCH program of Felsenstein's Phylogenetic Inference Package (PHYLIP) and the principal component projection to make an exhaustive examination of 13 amino acid exchange matrices. Here we apply a new artificial neural network (ANN)-based approach to examine these matrices and cluster the amino acids.

The ANN method exploited here is our recently developed Self-Organizing Tree Algorithm (SOTA; Dopazo and Carazo, 1997), which is based on both Kohonen's self-organizing mapping (Kohonen, 1990) and Fritzke's growing cell structure (Fritzke, 1994). In this work, we have used a modified form of SOTA, which we will refer to as SOTA/DIST, which was originally designed to cluster protein sequences, based on their distance matrix, although its use here has been to cluster amino acids based on an amino acid exchange matrix. Thirteen such scoring matrices were extracted from Johnson's collection of 15 amino acid exchange matrices (http://www.btk.utu.fi/molmol/matrices.html), including those of Dayhoff PAM250, Doolittle, Fitch, Gonnet, Grantham, Henikoff, Johnson, Jones, Levin, McLachlan, Miyata, Rao and Risler. L_1 distances were calculated between pairs of columns in a particular matrix. For example, the distance between alanine and cysteine scoring distributions of a matrix, X, is calculated as follows:

$$D_{ala,cys} = |X_{ala\to ala} - X_{cys\to ala}| + |X_{ala\to cys} - X_{cys\to cys}| + |X_{ala\to asp} - X_{cys\to asp}| + \ldots + |X_{ala\to tyr} - X_{cys\to tyr}|$$

We further cluster the 13 matrices with SOTA/DIST, to examine relationships among them. The Euclidean distance between every two matrices (X, Y) was calculated following the equation described by Johnson and Overington (1993):

$$D_{x,y} = [(X_{ala\to ala} - Y_{ala\to ala})^2 + (X_{ala\to cys} - Y_{ala\to cys})^2 + (X_{ala\to asp} - Y_{ala\to asp})^2 + \ldots + (X_{tyr\to trp} - Y_{tyr\to trp})^2 + (X_{tyr\to tyr} - Y_{tyr\to tyr})^2]^{1/2}$$

When running SOTA/DIST to cluster amino acids using each of the 13 scoring matrices, we reached algorithmic convergence in all cases. Five of the 13 dendrograms corresponding to these classifications are shown in Figure 1a–e (the other eight dendrograms are not shown). Results based on nine of the 13 matrices (those of Dayhoff, Doolittle, Gonnet, Grantham, Henikoff, Johnson, Jones, McLachlan and Miyata) show that the amino acids are grouped into two main clusters: small, polar and charged side chains {A, G, S, T, P, D, E, N, Q, H, K, R}, and hydrophobic side chains {C, I, V, L, M, F, Y, W}. Within the first cluster, small amino acids and charged amino acids/acid amides are separated into two subclusters. Within the second cluster, {I, L, M, V} and aromatic {F, W, Y} are two subclusters, while the location of C is changed variously. In most cases, C and W are much distant from other amino acids. These results correspond to the well-known groupings of amino acids: volume, hydrophobicity, charges (N-acid, H-basic), acid amide and aromatic groups. The grouping by SOTA/DIST based on Dayhoff PAM250 is perfectly consistent with the original classification by George *et al.* (1990) (Figure 1a), followed by groupings based on the matrices of Grantham, Henikoff, Johnson and Jones. Groupings based on the matrices of Doolittle, Gonnet and McLachlan exhibit similar overall characteristics of physicochemical relationships. The matrix of Levin, as well as the three matrices based on two-dimensional or three-dimensional structures as well as residue volumes (Miyata, Rao and Risler), only

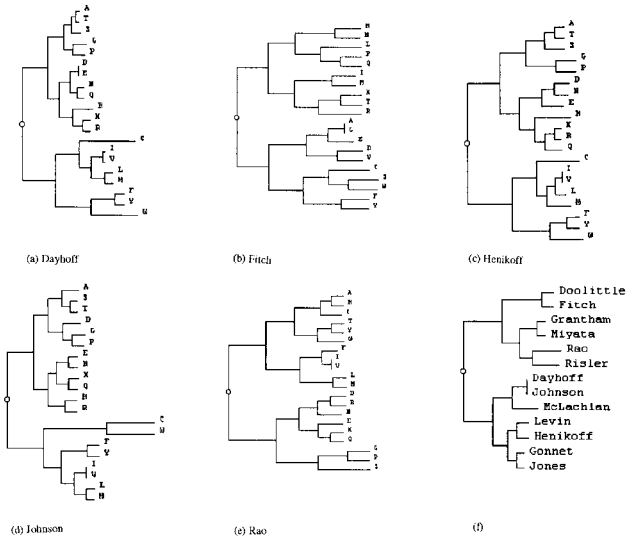

Fig. 1. (**a–e**) Dendrograms of the 20 amino acids constructed by the SOTA/DIST algorithm based on five amino acid scoring matrices. (**f**) Dendrogram of the 13 amino acid exchange matrices constructed by SOTA/DIST.

show parts of the characteristics mentioned above. The only exception to the general classification schema is the classification obtained using the matrix of Fitch, which does not show any noticeable physicochemical relationships of amino acids (Figure 1b). This last result is not too surprising, for the Fitch matrix is only derived from similarity of genetic encoding of amino acids. The same result was reached based on the matrix by Sammon mapping (Agrafiotis, 1997).

SOTA/DIST was then used to cluster the scoring matrices themselves. The result shows that the 13 matrices are clustered into several distinct groups that are consistent with the foundation on which they were based (Figure 1f). The matrices of Dayhoff, Johnson, McLachlan, Levin, Henikoff, Gonnet and Jones, which are all derived from sequence family (blocks) alignment and five of which are based on observed frequency of residue mutations, are in the same cluster. The matrix of Doolittle, which is based on both structure and genetic encoding, as well as the matrix of Fitch, based only on genetic encoding, are clustered together. The last result was also obtained when using the hierarchical clustering approach of the UPGMA (average linkage) method, but not when using the PHYLIP/KITSCH method (Johnson and Overington, 1993). The four matrices based on residue volumes, two-dimensional and three-dimensional structures (the Grantham, Miyata, Risler and Rao matrices) are all in a distinct cluster. A comparison of the results of grouping the 13 matrices previously performed by Johnson and Overington (1993) with principal components projection and hierarchical clustering by PHYLIP/KITSCH clearly indicates that the result by SOTA/DIST is quite similar to that obtained by principal components analysis rather than by hierarchical clustering.

Considering SOTA's performance on classifying amino acids and amino acid exchange matrices, we can conclude that the neural network approach used here is able to capture the essential features of a scoring matrix that corresponds with the physicochemical and structural properties of the amino acids, as well as to classify different scoring matrices according to the way in which they were derived. The relationship among amino acid properties is inherently non-linear and a neural network is very suitable for such a task and, in theory, it can grasp all of this kind of relationships. This is why SOTA can successfully classify the amino acids. We expect that the SOTA architecture can be used for a whole host of classification tasks that extend beyond sequence comparison, and is an appealing alternative to traditional clustering techniques when a complex and non-linear relationship among data is to be analysed.

Acknowledgements

We thank Mark Johnson for explaining the use of his collection of 15 amino acid exchange matrices. H.-C.Wang is supported by a CSIC fellowship for scholar exchange between Spain and China. Financial help from the Spanish CICYT through project number BIO95-0768 is greatly appreciated.

References

Agrafiotis,D.K. (1997) A new method for analyzing protein sequence relationships based on Sammon maps. *Protein Sci.*, **6**, 287–293.

Dickerson,R.E. and Geis,I. (1969) *The Structure and Action of Proteins*. Harper and Row, New York.

Doolittle,R.F. (1979) Protein evolution. In Neurath,H. and Hill,R.L. (eds), *The Proteins*. Academic Press, New York, Vol. 4, pp. 1–118.

Dopazo,J. and Carazo,J.M. (1997) Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol.*, **44**, 226–233.

Fritzke,B. (1994) Growing cell structures—a self-organizing network for unsupervised and supervised learning. *Neur. Net.*, **7**, 1141–1160.

George,D.G., Barker,W.G. and Hunt,L.T. (1990) Mutation data matrix and its uses. *Methods Enzymol.*, **183**, 333–351.

Johnson,M.S. and Overington,J.P. (1993) A structural basis for sequence comparisons—an evaluation of scoring methodologies. *J. Mol. Biol.*, **233**, 716–738.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Applic. Biosci.*, **8**, 275–282.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A mutation data matrix for transmembrane proteins. *FEBS Lett.*, **339**, 269–275.

Kohonen,T. (1990) The self-organizing map. *Proc. IEEE*, **78**, 1464–1480.

Sneath,P.H.A. (1966) Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.*, **12**, 157–195.

Taylor,W.R. (1986) Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.*, **188**, 233–258.

Taylor,W.R. and Jones,D.T. (1993) Deriving an amino acid distance matrix. *J. Theor. Biol.*, **164**, 65–83.