Theor Chem Acc (1999) 101:77–81
DOI 10.1007/s002149800m83

Theoretical
Chemistry Accounts

Regular article

Sequence classification of water channels and related proteins in view of functional predictions*

B. Tallur[1], **J. Nicolas**[1], **A. Froger**[2], **D. Thomas**[2], **C. Delamarche**[2]

[1] IRISA, Campus Universitaire de Beaulieu, Avenue de Général Leclerc, F-35042 Rennes Cedex, France
[2] CNRS UPRES-A 6026, Campus Universitaire de Beaulieu, Avenue de Général Leclerc, F-35042 Rennes Cedex, France

Received: 24 April 1998 / Accepted: 4 August 1998 / Published online: 2 November 1998

Abstract. We have worked with a classification method based upon a notion of probabilistic similarity or "likelihood of similarity" between aligned sequences. One important parameter, among others, affecting the sequence similarities and hence the classification results is the amino acid similarity matrix. We present a method for choosing the most adapted matrix to classify protein sequences. This method has been applied to the transmembrane channels of the major intrinsic protein (MIP) family. At present, two functional subgroups have been well characterized in this family: (1) specific water transport by the aquaporins and (2) small neutral solutes transport. The aim of the present study is to show the usefulness of the classification method in the prediction of sequence segments important for substrate selectivity. Moreover, we show that this method can also be used to predict the function of undetermined MIP proteins. The method could be applied to other protein families as well.

Key words: Hierarchical classification – Biological sequences – Major intrinsic protein family – Likelihood link analysis – Similarity matrix – Data mining

1 Introduction

The major intrinsic protein (MIP) family is an old family of transmembrane channels, including more than 150 members determined from bacteria, yeasts, plants, and animals [6, 14]. Less than 40 of these proteins have been functionally characterized to-date, exhibiting two major types of channel properties: specific water transport by the aquaporins (AQP) and small solute transport, such as glycerol by the glycerol facilitators (GLPF). The existence of aquaporins was suspected for a long time, but cloning and characterization of the first water channel is recent [16]. Aquaporins are characterized from bacteria to animals. At present, eight mammalian aquaporins have been identified (AQP1 through AQP8). All aquaporins are water specific, except for AQP3 and AQP7 which transport water and present a low permeability for small solutes such as glycerol and urea. Sequence alignments revealed that the GLPF and the aquaporins are related proteins. The GLPF channels exclude water but act as a selective pore for some uncharged molecules [7]. Actually, glycerol facilitators have been cloned from different bacteria and from yeasts. Aquaporins and glycerol facilitators are classified in the MIP family (PROSITE database, PS00221) with reference to its archetype MIP26, the major intrinsic protein expressed in lens fiber cells [5]. Members of the MIP family are about 260 residues long, and exhibit six transmembrane domains (Fig. 1) [6]. Highly conserved motifs, distributed throughout the sequence, constitute a signature of the MIP family, but are independent of functional properties. Extensive molecular studies on AQP1 led to a topological model [9], but these studies are both insufficient to locate precisely the aqueous pore and to point out residues implicated in the selectivity for water or glycerol. In order to discover amino acids which should be responsible for the functional properties of a protein, molecular biologists commonly use site-directed mutagenesis experiments. However, computational predictions are of particular relevance to avoid random targeting, which can be both laborious and expensive. Multiple sequence comparisons are classical tools for this purpose, and the information content of a multiple alignment can be extracted by different methods [1, 6, 12, 13]. Methods of sequence classification based on probabilistic similarities between sequences have been proposed and tested on several sets of multiple aligned, as well as unaligned, sequences [11, 19], and it was clearly established that one of the major parameters affecting the classification results is the amino acid similarity matrix used in the evaluation of similarities between sequence pairs. The biologists have

*Contribution to the proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambery, France

Correspondence to: B. Tallur
Tel.: +33-2-99847300, Fax: +33-2-99847171,
e-mail: tallur@irisa.fr

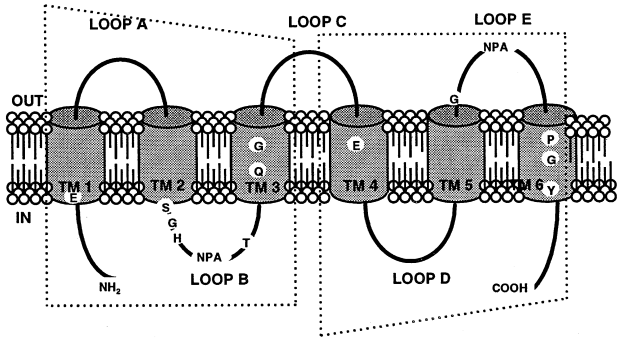


Fig. 1. Predicted membrane topology of a monomer of the MIP family protein, showing the six transmembrane domains with five connecting loops. The predicted ancestral tandem duplication event is indicated with *dotted boxes* [16]. The location of highly conserved residues is shown

a very large choice of similarity matrices such as PAM250 [3], BLOSUM62 [8], Risler's matrix [17], and many others based on different molecular properties. This paper proposes an efficient method for choosing the most adapted similarity matrix to classify protein sequences. It has been illustrated with an application to the MIP family.

2 Methodology

2.1 Selection of protein sequences

A set of 38 MIP protein sequences was extracted from GENBANK and PROSITE databases. As mammalian and plant aquaporins are over-represented in databases and could influence the results of sequence alignments, we selected only a few members of these two groups of organisms, rather than retrieving all members of the MIP family. The MIP sequences were classified into two functional subgroups: specificity for water transport (AQP subgroup) or solute transport (GLPF subgroup, including glycerol transport and mixed channels). AQP subgroup: M84344, P30302, P43285, and U39485 (*A. thaliana*), X97159 (*C. viridis*) U38664 (*E. coli*), Q39957 and X95952 (*H. annuus*), U51638 (*H. irritans*), D31846, D63412, M77829, P55064, Q13520, and U34846 (*H. sapiens*), L36095, Q40260, and Q40266 (*M. crystallinum*), L24754 (*R. esculenta*), D13906, U14007, U16245, and X70257 (*R. norvegicus*). GLPF subgroup: P18156 (*B. subtilis*), M55990 (*E. coli*), P44826 (*H. influenzae*), AB006190 and D25280 (*H. sapiens*), M58315 (*L. lactis*), U49666 (*P. aeruginosa*), L35108 and P56403 (*R. norvegicus*), P23900 (*S. cerevisae*), P37451 (*S. thyphimurium*), P31140 (*S. flexneri*), U12567 (*S. pneumoniae*), Q21473 and Q21159 (*C. elegans*).

2.2 Sequence alignment

The PILEUP program (version 8) of the GCG package [4] was used to align 32 among the 38 MIP protein sequences (19 AQP and 13 GLPF). This generated a final alignment with 422 positions when excluding the long N- and C-terminal unusual segments of the yeast aquaporin P23900. In a second step, we used the TMAP program which efficiently predicts transmembrane segments from protein sequence alignments [15]. From these alignments we extracted the predicted blocks (LOOPA through LOOPE and TM1 through TM6 shown in Fig. 1), but excluded the NH_2 and COOH termini blocks which contained a large number of gaps introduced during the alignment procedure. Each block line corresponds to one sequence and is composed of a string of letters belonging to an alphabet of 20 letters (representing amino acids), augmented with the gap symbol "–". Thus, the alphabet considered in this study is $\mathscr{L} = \{ACDEFGHIKLMNPQRSTVWY-\}$.

2.3 Sequence classification

Though there are a large number of sequence classification methods available for the biologists, we have worked with the hierarchical classification based on the link likelihood analysis (LLA) developed by Lerman [10]. One of the main reasons for this choice is that we have devised (in our previous work) efficient measures of similarity between protein sequences that are suitable for the LLA methodology, and found it efficient in various applications. The important features that distinguish the LLA method from the rest are: (1) the probabilistic nature of the similarity index over the set of objects to be classified (e.g. set of sequences); (2) the aggregation criterion – for building the hierarchical tree by joining the most similar classes at each level – that takes into account the specific nature of the similarity index. A brief description of LLA is attempted in the following subsections.

2.3.1 Significant windows approach

The "significant windows" approach is used for computing the similarities between sequences. This approach makes use of the matrix of similarities between amino acids and the final results are highly dependent on the matrix chosen. Very frequently, biologists run one of the numerous multiple alignment programs on their sequences in order to detect the highly preserved sites and the results produced vary from one algorithm to another. The methods which make use of the similarity between sequences based on the site-wise comparison of letters are clearly sensitive to the possible alignment errors. The significant windows approach – described in [11] for the aligned set of sequences and in [19] for the unaligned sequences – tends to be less affected by such errors. Another important objective of this approach is to select only the relevant or "significant" information concerning the problem. The overall similarity between a given sequence pair is computed from the set of similarities between the significant windows. (See Refs. [11] and [19] for the method of selecting the set of "significant" windows.)

2.3.2 Window similarity for a sequence pair

A window of size l of a sequence is a sub-sequence containing l letters. The ith window of size l starts at position i and terminates at position $(i+l-1)$. Let us suppose that the common length of the (aligned) sequences is L; by sliding the l-size window along the sequences, one gets $(L-l+1)$ windows. Let a_i^j denote the letter at position i in sequence j. To fix the ideas, let us consider the set $\mathscr{O}$ of all sequences. Consider the following sequence pair $(o_j, o_{j'})$:

$$\text{sequence } o_j: \quad a_1^j, a_2^j, \ldots, a_i^j, \ldots, a_{i+l-1}^j, a_{i+l}^j, \ldots, a_L^j$$

$$\text{sequence } o_{j'}: \quad a_1^{j'}, a_2^{j'}, \ldots, a_i^{j'}, \ldots, a_{i+l-1}^{j'}, a_{i+l}^{j'}, \ldots, a_L^{j'}$$

Let W_{ij} and $W_{ij'}$ denote respectively the ith windows of o_j and $o_{j'}$. The ith window similarity $S_{jj'}^i$ for the comparison of the sequence pair $(o_j, o_{j'})$ is the sum of the standardized scores of the matrix D^s – obtained by standardizing one of those amino acid similarity matrices with respect to its mean and standard deviation – corresponding to the l letter pairs of the window:

$$S_{jj'}^i = S(W_{ij}, W_{ij'}) = \sum_{k=i}^{i+l-1} D^s(a_k^j, a_k^{j'}) \tag{1}$$

Thus, obviously, the similarity of any sequence pair completely depends on the particular matrix chosen. Let us denote the overall similarity of the sequence pair $(o_j, o_{j'})$ by $S(o_j, o_{j'})$. They are finally standardized with respect to the empirical distribution over the set of all sequence pairs $\mathscr{P}_2(\mathscr{O})$, and the standardized similarity index is denoted by $Q_s(o_j, o_{j'})$.

2.3.3 Aggregation criterion used in LLA method

The basic data required by the LLA method of hierarchical classification is the matrix of probabilistic similarities between the sequence pairs, given by the equation

$$P(o_j, o_{j'}) = \Phi[Q_s(o_j, o_{j'})], \quad \text{for } (o_j, o_{j'}) \in \mathscr{P}_2(\mathscr{O}) \tag{2}$$

where Φ denotes the standard normal distribution function and Q_s is the standardized similarity of the sequence pair (described in the previous paragraph). The algorithm builds a classification tree iteratively, by joining at each step the two (or more in case of ties) most similar sequences or classes of sequences until all clusters are merged together. The aggregation criterion that is maximized at each step or "level" of the algorithm is expressed as a similarity measure between clusters. Suppose that C and D are two arbitrary disjoint subsets (or clusters) of $\mathscr{O}$ comprising respectively r and s elements. Then a family of aggregation criteria of the "maximal link likelihood" is defined by the following measure of similarity between C and D:

$$LL_\gamma(C, D) = [\max\{P(c, d) : (c, d) \in C \times D\}]^{(rs)^\gamma} \quad 0 \le \gamma \le 1 \tag{3}$$

In case of our data sets, $\gamma = 0.5$ was found to yield the best results.

2.4 Similarity matrix selection

The underlying principle of our method for selecting the most relevant amino acid similarity matrix among all is simple: it is based on an indicator of the quality of a given partition – obtained for instance at a particular level of a hierarchical classification method – of the set $\mathscr{O}$ of sequences under study, expressed as a function of the similarity measure over $\mathscr{P}_2(\mathscr{O})$. The LLA method provides us with one such indicator called "global index" associated with each level of the hierarchical classification tree.

2.4.1 Indicator of the quality of a partition.

Let us consider the partition of $\mathscr{O}$ into k classes:

$$\pi = \{C_1, C_2, \ldots, C_k\} \ . \tag{4}$$

Let ω_m denote the pre-ordering over the set of sequence pairs induced by the similarities between sequences resulting from a given amino-acid similarity matrix m. The pre-ordering ω_m is mathematically represented by its graph $\mathrm{Gr}(\omega_m)$ where

$$\mathrm{Gr}(\omega_m) = \{(p, q) \ : \ p \prec q, \ p \in \mathscr{P}_2(\mathscr{O}), \ q \in \mathscr{P}_2(\mathscr{O})\} \tag{5}$$

and $p \prec q$ means sequence pair p precedes q for the order relation induced by the similarity measure (i.e. the similarity of the pair p is greater than that of the pair q). Similarly, the partition π is represented by the following two sets:

$$V = \{(o_i, o_j) : o_i \in C_l \text{ and } o_j \in C_l, \ l = 1, \ldots, k\} \tag{6}$$

is the set of all sequence pairs both elements of which belong to the same class and

$$W = \{(o_i, o_j) \ : \ o_i \in C_l \text{ and } s_j \in C_m, \ l \neq m\} \tag{7}$$

the set of sequence pairs whose elements lie in distinct classes.

The "global index" associated with the partition is expressed as a similarity index between the preordering ω_m, on the one hand, and the partition π represented by the sets V and W, on the other. The raw index of similarity between ω and π is defined as

$$s_{\omega\pi} = \mathrm{Card}(\mathrm{Gr}(\omega) \cap (V \times W)) \tag{8}$$

where Card is the cardinal and $\times$ denotes the Cartesian product. The final index (or the global index) $\mathrm{globstat}_{m,\pi}$ is obtained by standardizing the above raw index with respect to its mean and standard deviation.

2.4.2 Matrix selection method.

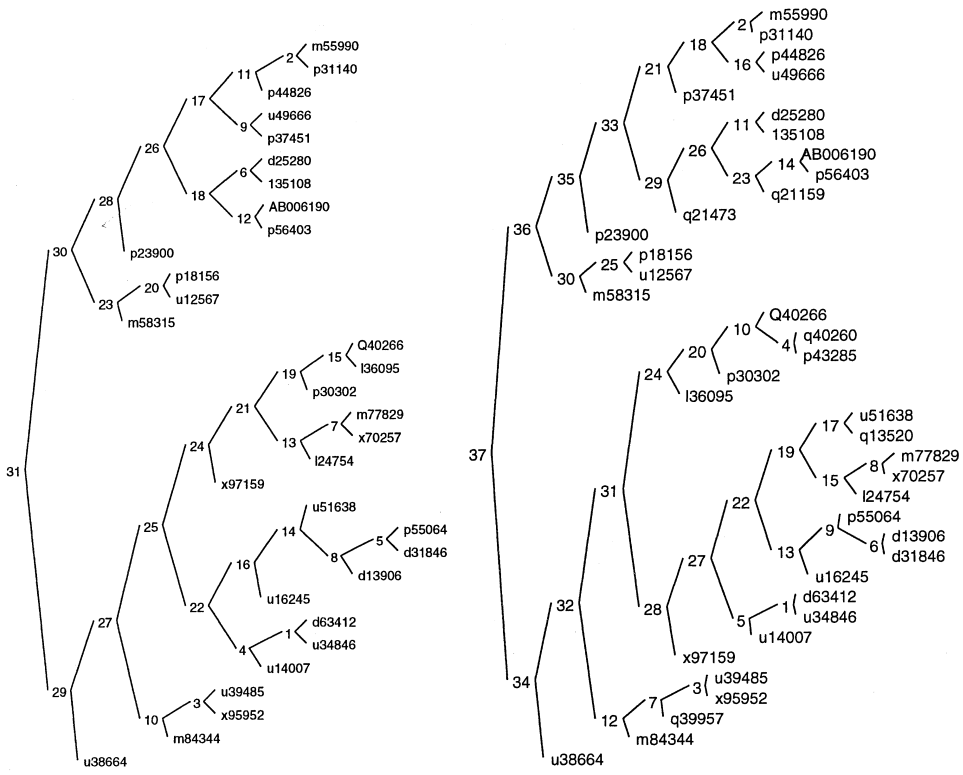
Let us consider a special case where the set $\mathscr{O}$ is partitioned into two classes $\mathscr{O}_{\mathrm{aqp}}$ and $\mathscr{O}_{\mathrm{glpf}}$ comprising respectively the known AQP and GLPF protein sequences, i.e. $\pi = \{\mathscr{O}_{\mathrm{aqp}}, \mathscr{O}_{\mathrm{glpf}}\}$. Suppose $\mathscr{M}$ is the set of amino acid similarity matrices to choose from and π is the partition of sequence set into two classes containing respectively the sequences whose function is determined as AQP and GLPF sequences. The idea is to pick that matrix $m \in \mathscr{M}$ which induces over the set $\mathscr{P}_2(\mathscr{O})$ the preorder that is most consistent with the partition of $\mathscr{O}$ into two known classes of sequences. Therefore we consider as the best matrix for the prediction of MIP protein function the one that maximizes – over the set of possible matrices m – the global statistic $\mathrm{globstat}_{m,\pi}$ comparing ω_m with the partition of π.

3 Results

Aquaporins and glycerol facilitators belong to the same old family of channel proteins, but at present it is impossible to understand why some highly similar protein sequences have so dissimilar functional properties. Recently, we reported a sequence comparison study of the MIP family proteins based on similarity profile analysis and multivariate statistical analysis [6]. Five key residues were predicted to play a role in the structural/functional properties of the MIP proteins, but, while these residues are important, we do not know whether they are sufficient to explain the drastic functional differences between the aquaporins and the glycerol facilitators. Transmembrane segments and loops represent important and distinct parts in transmembrane proteins. Therefore, we selected these regions to highlight segments which contribute most to functional differences in the MIP family. The predictive method described in the previous section was applied to each segment: LOOPA through LOOPE and TM1 through TM6. Three similarity matrices were compared, namely, Dayhoff's PAM250 matrix, BLOSUM62, and Risler's matrix. For the LOOPC region, Risler's matrix was most appropriate (i.e. produced the highest $\mathrm{globstat}_{m,\pi}$). The use of the jackknife procedure showed this result to be stable.

The hierarchical classification of LOOPC segments was then performed using the most appropriate similarity matrix chosen as above. The alignment of 32 proteins with a known function was used in the first phase. Our assumption is that segments which give a perfect partition into two classes corresponding to AQP and GLPF proteins are likely to be important for the substrate selectivity of these proteins. Such a partition was obtained for LOOPC and is presented as an example in this study (Fig. 2a). A close inspection of each branch of the hierarchical classification tree should highlight subtle evolutionary events and functional variations linked to the segment. In that sense, our method is similar to the evolutionary trace method described by Lichtarge et al. [12]. For example, a subclass of GLPF sequences P18156, U12567, and M58315 was found in all classification trees produced by using any of the similarity matrices. However, as suggested by a recent study, these sequences are reported as belonging to two distinct evolutive sub-families [14]. According to our analysis the grouping of these proteins may be explained by the fact that they exhibit very close functional properties: our GLPF subgroup includes channels for glycerol transport and also those for the transport of other small solutes such as propanediol or antimonite [2, 18]. This result is in accordance with our previous results obtained with another method, suggesting that LOOPC should be implicated in solute transport properties [6].

In the second phase, the classification method using Risler's matrix was applied to the set of 38 sequences comprising the 32 previous sequences to which were



(a) Classification of 32 LOOPC segments

(b) Classification of 38 LOOPC segments

Fig. 2. Classification trees of LOOPC sequence segments with the significant windows approach (Risler matrix, window size = 6). Each sequence is identified by its accession number in the databases

added 6 test sequences with a known function (4 AQP and 2 GLPF). The classification tree (Fig. 2b) has 37 levels and each of the two classes of the partition produced at level 36 may be distinctly identified to one functional group. Moreover, the 6 test sequences are correctly classified. The quality of the prediction was evaluated by the jackknife technique. The repeated trials, in which one sequence was removed at a time from the set, yielded a 100% correct prediction for the 6 test sequences.

At present, the MIP family includes more than 150 sequences but only a few of them have been biochemically characterized and there is no simple and straightforward method to study solute transport through specific channels. Consequently, our classification method could be helpful to predict the functional properties of the proteins which are not yet experimentally determined. One may reasonably suppose that an "undetermined" protein found in one branch of the tree has functional relations with that branch.

4 Conclusion

A hierarchical classification method based on significant windows approach has been successfully applied to the two functional classes of the MIP family. More generally, the method can be applied to predict any number of classes and it proves to be a precious data mining tool for biologists since it can be applied for tuning automatically other parameters such as window size and significance level that also affect the classification.

References

1. Andrade MA, Casari, G, Sanders C, Valencia A (1997) Biol Cybern 76:441
2. Chen P, Anderson DI, Roth JR (1994) J Bacteriol 176:5474
3. Dayhoff MO, Barker WC, Hunt LT (1983) Methods Enzymol 91:524
4. Devereux J, Haeberli P, Smithies O (1984) Nucleic Acids Res 12:387
5. Gorin MB, Yancey SB, Cline J, Revel JP, Horwitz J (1984) Cell 39:49
6. Froger A, Tallur B, Thomas D, Delamarche C (1998) Protein Sci 7:1458
7. Heller KB, Lin ECC, Wilson TH (1980) J Bacteriol 144:274
8. Henikoff S, Henikoff JG (1992) Proc Natl Acad Sci USA 89:10915
9. Jung JS, Preston GM, Smith BL, Guggino WB, Agre P (1994) J Biol Chem 269:14648
10. Lerman IC (1991) App Stochastic Data Anal 7:63
11. Lerman IC, Nicolas J, Tallur B, Peter P (1994) In: Diday E, Lechevallier Y, Schader M, Bertrand P, Burtschy B (eds) New approaches in classification and data analysis. Springer, Berlin Heidelberg New York, pp 370–377
12. Lichtarge O, Bourne HR, Cohen FE (1996) J Mol Biol 257:342
13. Livingstone CD, Barton GJ (1993) Comput Appl Biosci 9:745
14. Park JH, Saier HM Jr (1996) J Membr Biol 153:171
15. Persson B, Argos P (1994) J Mol Biol 237:182

16. Preston GM, Carrol TP, Guggino WB, Agre P (1992) J Biol Chem 268:17
17. Risler JL, Delorme MO, Delacroix H, Hénault A (1988) J Mol Biol 204:1019
18. Sandres OI, Rensing C, Kuroda M, Mitra B, Rosen BP (1997) J Bacteriol 179:3365
19. Tallur B, Nicolas J (1997) In: Hayashi C, Ohsumi N, Yajima K, Tanaka Y, Bock HH, Baba Y (eds) Data science, classification and related methods. Springer, Tokyo Berlin Heidelberg New York, pp 758–765