Color and Graphic Display (CGD): Programs for Multiple Sequence Alignment Analysis in Spreadsheet Software

BioTechniques 29:100-107 (July 2000)

Christian Delamarche

Université de Rennes 1, Rennes, France

ABSTRACT

Interpretation of multiple sequence alignments is of major interest for the prediction of functional and structural domains in proteins or for the organization of related sequences in families and subfamilies. However, a necessity for the bench scientist is the use of outstanding programs in a friendly computing environment. This paper describes Color and Graphic Display (CGD), a set of modules that runs as part of the Microsoft[®] Excel[®] spreadsheet to color and analyze multiple sequence alignments. Discussed here are the main functions of CGD and the use of the program to highlight residues of importance in a water channel family. Although CGD was created for protein sequences, most of the modules are compatible with DNA sequences.

INTRODUCTION

Many commercial or freeware programs for alignment display and analysis are available (http://mercury.ebi.ac.uk/bio-cat/Alignment_editing_and_display.html) (11,20). Unfortunately, many of these programs are not useful for biology students or researchers because they offer a limited choice of methods and optional parameters (input and output formats, graphical types), or they may require hours of practice in unfamiliar environments, for example, using the UNIX[®] operating system or Adobe[®] PostScript[®] files.

Becoming familiar with a given piece of software can be time consuming. Thus, a non-expert in sequence analysis would prefer to use multipurpose software. Spreadsheet software is useful, widely used by biologists and provides many statistical tools, sorting functions and chart types. The spreadsheet software, Microsoft[®] Excel[®], has been used to calculate enzyme kinetic parameters (18), to color amino acids (14) and to predict structural domains in proteins (2,6,13,30). This software has been used to develop Color and Graphic Display (CGD), a package of 14 modules dedicated to multiple sequence alignment analysis.

For this paper, CGD was applied to the prediction of biological data in major intrinsic protein (MIP) channels, proteins that mediate flows of water or solutes (glycerol and oth-

er small uncharged molecules) across biological membranes (10). A topological model of an MIP channel consists of a homotetramer, in which each monomer is a pore and composed of six transmembrane α -helices connected by five loops. The internal loop B and the external loop E protrude into the pore, resulting in an hourglass shape (1,19). In this paper, sequence analysis is used to predict residues implicated in the selectivity of solutes by MIP channels.

SYSTEMS AND METHODS

Overview of CGD

The programs are written in Visual Basic[®] language (Microsoft). The present version of CGD (CGD.xla) has been tested on IBM PC-compatible computers with Excel 5.0, 95 and 97. (The extra-module logo.xla is not supported by Excel 95.) CGD does not create the alignments but is compatible with the widely used programs PILEUP, CLUSTAL, TREEALIGN, MULTALIGN, MAP, PIMA and BLOCKS (5).

First, the user copies the multiple alignment from a common text editor or from an HTML output file and pastes it into a cell of an Excel worksheet. When running the first module of CGD, the letters (amino acids or bases) are automatically redistributed into multiple cells. CGD uses conventional Excel dialog boxes such as color palette or format cell. Specific dialog boxes are also displayed in order to choose between algorithms or to classify amino acids into groups. In modules using computations, the results are presented in tables. The data can be plotted using the Chart Wizard of Excel, or they can be modified by applying user-defined equations.

The main modules available in CGD are briefly described here. CGD can be freely obtained from the author (cde-lam@univ-rennes1.fr) or from the supplementary materials link at www.BioTechniques.com.

The MIP Family

A set of 42 MIP protein sequences was extracted from GenBank[®], EMBL, TREMBL and SWISSPROT databases, which can be found at <http://www.ncbi.nlm.nih.gov/> or <http://www.ebi.ac.uk/embl/> or <http://www.epasy.ch/sprot/>, respectively. The sequences were partitioned into two functional subfamilies, one for water transport (AQP, aquaporin subgroup) and the other for solute transport (GlpF, glycerol

facilitators and mixed channels). AQP subgroup: Y07625 (NLM1), P93004 (PIP3), M84344 (TIPG), P43285 (WC1A), X68293 (WC1B), X69294 (WC1C), P43286 (WC2A), P43287 (WC2B), P30302 (WC2C) and U39485 (TIPD) from *A. thaliana*; X97159 (AQPcic) from *C. viridis*; U38664 (AqpZ) from *E. coli*; Q39935, Q39957 and Q39958 from *H. annuus*; M77829 (AQP1), D31846 (AQP2), D63412 (AQP4), P55064 (AQP5), Q13520 (AQP6) and P30301 (MIP26) from *H. sapiens*; L36095 (MIPA), Q40266 (MIPB) and Q40260 (MIPE) from *M. crystallinum*; L24754 (AQPA) from *R. esculenta*; X70257 (AQP1), D13906 (AQP2), U14007 (AQP4), U16245 (AQP5); and P56405 (AQP8) from *R. norvegicus*.

GlpF subgroup: P18156 (GlpF) from *B. subtilis*; M55990 (GlpF) from *E. coli*; P44826 from *H. influenzae*; D25280 (AQP3) and AB006190 (AQP7L) from *H. sapiens*; M58315 (YDP1) from *L. lactis*; U49666 from *P. aeruginosa*; L35108 (AQP3), P56403 (AQP7) and AF016406 (AQP9) from *R. norvegicus*; P23900 (FPS1) from *S. cerevisiae*; and P37451 (PDFU) from *S. thymophilum*.

The sequences were aligned with three programs: PILEUP within the GCG package (8), CLUSTALW 1.7 (31) and PRO-TOMAT (17). Each time, the global alignment of the 42 se-

quences was separated (using Excel's cut-and-paste function) into two distinct subalignments that correspond to sequences of the two functional subfamilies (30 AQP, 12 GlpF). Then, both the global alignment and each subalignment were analyzed with the CGD tools. Key positions were retrieved from regions giving consistent results with the different alignment programs.

RESULTS AND DISCUSSION

Coloring

Coloring identical or similar residues in multiple alignments is useful to predict transmembrane regions, derive consensus sequences, find periodic repeats of patterns or simply to compare a new sequence with other members of the family. In CGD, there are two ways to color the alignments. The first way colors residues that are identical or similar in all sequences; the second way colors residues using any reference sequence selected by the user. A dialog box is displayed to assign amino acids to groups of similarity. Cell patterns and letter fonts are selected from the color palette of Excel (Figure 1).

Figure 1. Three examples of multiple alignment presentations with CGD. (1) Similarities with sequence P43287 according to structural properties of the amino acids, (A, C, G, P, S, T, W, Y), (R, N, D, Q, E, H, K), (I, L, M, F, V). Numbers typed in the first block of column B permit the manual indexing of each sequence for computing the residue frequencies (see the weighting method). (2) Identities with sequence P55064. (3) Similarities with sequence P55064 according to chemical properties of the amino acids, (D, E), (H, R, K), (C, S, T), (N, Q), (M), (A, I, L, V), (F, W, Y), (P), (G).

Hydrophobicity and Similarity Plots

Hydrophobicity profiles are popular tools for predicting the topology of membrane proteins directly from their sequence. CGD proposes three scales of hydrophobicity (9,22,26) and two options. The first option assigns the hydrophathy values to a single sequence selected by the user, a method independent of the alignment. In the second option, CGD averages the individual hydrophathy indices at each position of the alignment. This method, based on the information given by the alignment, is more accurate because there is more information in a set of related sequences than in a single sequence (25).

One module of CGD calculates an average similarity score at each position along the multiple alignments. The score is the arithmetic mean of the distance values between all amino acid pairs. In an alignment of n sequences, the number of pairwise distances is as follows:

$$\binom{n}{2} = \frac{n!}{2(n-2)!}$$

and the distance values are taken from one of these three distance matrices: Blosum62 (15), Gribskov (12) and PAM250 (7).

Using the table of results created by CGD, hydrophobicity and similarity plots can be easily constructed for any window size, for the complete alignment or for some part of it. Figure 2

superimposes the average hydrophobicity and similarity plots for the alignment of 42 MIP sequences. This permits a rapid visualization of the alignment. The figure shows that loop B and E are highly conserved in the MIP family. Low values on the similarity plot indicate that these regions are poorly conserved. Negative values on the hydrophobicity plot correspond to extra amino acids (insertions) that characterized the GlpF subgroup in loop D and at the end of loop E (10,21).

Information Quantification

A module of CGD determines the frequency of occurrence $f(x, L)$ of each symbol x at each position L of the alignment. From this frequency table, it is easy to calculate the information content along the alignment, with the help of the worksheet functions of Excel. A nice method, based on the Shannon information theory (29), has been introduced by Schneider (28):

$$I_{\text{seq}}(L) = \log_2 S - H(L) - e(n) \text{ where } H(L) = -\sum_{x=A}^S f(x, L) \log_2[f(x, L)]$$

$I_{\text{seq}}(L)$, measured in bits, is the information at position L , S is the number of symbols (4 for nucleotides or 20 for amino acids), $\log_2 S$ is the maximum value of uncertainty, n is the number of sequences, $H(L)$ is the uncertainty at position L and $e(n) = (S - 1)/(2 n \ln 2)$ is a correction factor for finite-size

Figure 2. Average similarity (continuous curve) and hydrophathy (dotted curve) plots of the MIP family proteins. CGD program was used with the similarity matrix of Gribskov (12) and hydrophathy scale of Kyte and Doolittle (22), with a window size of 19 residues. For convenience, the N- and C-termini are not shown.

Table 1. Sequence Specificity of Each Functional Subgroup of the MIP Family

pos.	loc.	AQP subgroup										GlpF subgroup										r
		DE	HRK	CST	NQ	M	AILV	FWY	P	G	DE	HRK	CST	NQ	M	AILV	FWY	P	G			
259	T 1	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.00		
267	T 1	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	13.3	86.7	0.0	0.0	0.0	0.99			
326	T 2	0.0	0.0	0.0	0.0	0.0	98.2	0.0	1.8	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	1.00			
328	T 2	0.0	0.0	0.0	0.0	3.6	83.9	12.5	0.0	0.0	0.0	0.0	0.0	0.0	88.9	11.1	0.0	0.0	1.00			
334	T 2	0.0	0.0	1.8	0.0	0.0	98.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	1.00			
346	Lp B	0.0	0.0	1.8	0.0	0.0	98.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	1.00			
347	Lp B	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.00			
348	Lp B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	5.6	0.0	0.0	0.0	0.0	94.4	1.00			
350	Lp B	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	94.4	0.0	5.6	0.0	0.0	0.0	1.00			
352	Lp B	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	1.00			
353	Lp B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	6.7	0.0	93.3	0.0	1.00			
355	Lp B	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	94.4	5.6	0.0	0.0	1.00			
356	Lp B	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	1.00			
370	T 3	0.0	0.0	3.6	0.0	0.0	90.1	0.0	0.0	6.3	0.0	0.0	0.0	0.0	80.0	20.0	0.0	0.0	0.96			
373	T 3	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	1.00			
375	T 3	0.0	0.0	1.8	0.0	0.0	98.2	0.0	0.0	0.0	0.0	0.0	5.6	0.0	0.0	87.8	0.0	0.0	1.00			
377	T 3	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	1.00			
379	T 3	0.0	0.0	0.0	0.0	0.0	98.2	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	1.00			
380	T 3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	1.00			
387	T 3	0.0	0.0	1.3	0.0	0.0	93.8	5.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	80.0	0.0	0.0	0.97			
388	T 3	0.0	1.3	0.0	0.0	0.0	98.7	0.0	0.0	0.0	0.0	0.0	6.7	0.0	0.0	93.3	0.0	0.0	1.00			
392	Lp C	0.0	0.0	60.8	10.0	4.3	23.6	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	-0.18			
436	T 4	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	80.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.97			
505	T 5	0.0	0.0	0.0	0.0	0.0	98.2	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	1.00			
509	T 5	0.0	90.7	0.0	0.0	0.0	9.3	0.0	0.0	0.0	0.0	0.0	6.7	33.3	0.0	5.6	0.0	54.4	-0.20			
516	Lp E	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	1.00			
519	Lp E	0.0	0.0	62.0	0.0	0.0	0.0	0.0	0.0	38.0	0.0	0.0	0.0	0.0	94.4	0.0	0.0	5.6	-0.14			
521	Lp E	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.00			
523	Lp E	0.0	0.0	0.0	0.0	0.0	98.7	0.0	0.0	1.3	0.0	0.0	13.3	0.0	0.0	86.7	0.0	0.0	0.99			
524	Lp E	0.0	98.7	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	1.00			
525	Lp E	0.0	0.0	96.9	0.0	0.0	3.1	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.11			
529	Lp E	0.0	0.0	2.5	0.0	0.0	97.5	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.11			
555	T 6	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	-0.11			
556	T 6	0.0	0.0	0.0	0.0	0.0	1.3	98.7	0.0	0.0	0.0	0.0	0.0	38.9	61.1	0.0	0.0	0.0	-0.15			
559	T 6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	1.00			
561	T 6	0.0	0.0	0.0	0.0	1.8	98.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	1.00			
562	T 6	0.0	0.0	0.0	0.0	0.0	1.8	0.0	0.0	98.2	0.0	0.0	0.0	0.0	5.6	0.0	0.0	94.4	1.00			
570	T 6	0.0	12.5	0.0	0.0	0.0	1.8	85.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.99			

Column 1, position in the global multiple alignment; column 2, probable localization in a transmembrane or a loop region; columns 3–20, frequency of each amino acid group, expressed in percentage; column 21, correlation coefficient. Shaded lines point out amino acid groups that could modulate the selectivity of MIP channels.

samples. A sequence logo is a graphical representation of $I_{\text{seq}}(L)$, in which the letters of the sequences are stacked on top of each other, at each position of the alignment (27), as follows:

$$\text{Height of letter } x \text{ at position } L = f(x, L) I_{\text{seq}}(L)$$

The size of each letter grows with the increasing frequency of the letter, and the stack height of letters represents the total information at that position. CGD includes an Excel worksheet, dedicated to protein logos (only Excel 97 and further versions). One modification in the computing of $I_{\text{seq}}(L)$, was to add the information given by gaps that are introduced in sequence alignments. Although gaps do not really exist, their

presence in multiple alignments can reveal insertion/deletion events important for the structure or the function of the molecule. Therefore, in CGD, the symbol S represents 21 equally probable symbols (20 amino acids and gaps). This method reduces the stack height at positions containing many gaps. Logos are more subtle than consensus sequences because they associate a mathematical formulation of sequence conservation with a colored representation of amino acid classes. For example, the height of Q at position 350 of the MIP logo (Figure 3) indicates a low contribution of glutamine in that position of the pattern. Indeed, only one MIP sequence matches with the pattern Q-x-N-P-A-V-T.

Sequence Weighting

In a multiple alignment, the results of a counting method—and subsequently, the predictions based on it—can be biased by redundancies in the source of sequences. For example, the information contained in the orthologous sequences will dominate if each sequence is considered independently in a multiple alignment composed of many sequences from one evolutionary group and few sequences from another group. It is possible to improve the performance of a counting method in assigning weights to sequences (3,16). CGD provides a simple and original method to represent the heterogeneity in a set of aligned sequences. The user indexes the sequences into subclasses (evolutionary kingdoms, structural/functional particularities, clusters in a dendrogram and so forth). All sequences with the same attribute are considered as one. The result is that, when counting one residue at any position, CGD divides it by the number of sequences included in the corresponding subclass. In the example of Figure 1.1, if the counting method is applied without correction, the results for the first and second column of the alignment are respectively 5 M, 1 gap and then 3 A, 1 K, 1 D, 1 gap (total of each column = 6). When the sequences are classified in three groups as indicated in column B, the new results of residue counting are respectively 2.5 M, 0.5 gap and then 1.83 A, 0.33 D, 0.33 K, 0.5 gap (total of each column = 3).

Filtering Methods and Statistical Analysis

The statistical methods and database functionality of Excel can be applied directly to the results generated by CGD. An example is shown in Table 1. First, the weighting method described earlier was applied to classify MIP sequences into six kingdoms (mammalian, amphibian, insect, plant, yeast and bacteria) to calculate the “corrected frequency” of each residue at each position of the AQP and GlpF subalignments. Then, the amino acids were grouped and the advanced filter command of Excel was used to select positions in the multiple alignment that follow given criteria. A position was considered prominent in the MIP family when the conservation of one amino acid group was $\geq 80\%$ in the two subfamilies, and a position was selective when the percentage of one amino acid group was $\geq 80\%$ in one subfamily and $\leq 20\%$ in the other (grey lines in Table 1).

The Pearson's correlation coefficient method, directly available in Excel, permits a statistical comparison of the distribution of the amino acid groups for the selected positions. The r values range from -1 to +1. Dissimilar positions between the subgroups have a negative r score (the groups are mutually determined, but they vary in the opposite sense). From Table 1, it is easy to describe the two major functional types of MIP channels from a limited number of key positions (gray areas). According to these predictions, we have recently shown that posi-

Figure 3. Logo obtained from the 42 MIP sequence alignment in region of loop B. Colors are assigned according to the Blockmaker criteria (17): red for negative (D, E), blue for positive (H, K, R), light grey for polar uncharged (C, S, T), green for amide (N, Q), yellow for methionine (M), black for hydrophobic (A, I, L, V), orange for aromatic (F, W, Y), purple for proline (P) and grey for glycine (G).

tions 555 and 556 (Table 1) participate in the selection of the transported molecule, water or glycerol (23,24). From prominent positions in the multiple alignment (table 1), it is also possible to define a new concept of signature sequence in which amino acids in brackets have common properties. For example, the signature [CST]-G-x-[HRK]-x-[NQ]-P-x-[AILV]-[CST], defined from positions 347–356, can be translated in a natural language: polar uncharged, glycine, any, positive, any, amide, proline, any, hydrophobic, polar uncharged. Such a description can help the structural interpretations. The new signature retrieves 203 MIP sequences in the SWISS-PROT and TREMBL databases at <http://www.expasy.ch/tools/scnpsit2.html>. The official signature of the MIP family, PS00221 in the PROSITE database (4), which is [HNQA]-x-N-P-[STA]-[LIVMF]-[ST]-[LIVMF]-[GSTAFY], retrieves 214 sequences. However, this sequence signature is less informative because it is only a consensus with no meaning in terms of the conservation of biological properties along the sequence.

CONCLUSION

CGD, software for multiple sequence alignment analysis, runs on personal computers equipped with Excel. The methods are easy to understand, and at any time, the user can follow and access the processing of its data. Many statistical methods are available in Excel, and users can select the most versatile presentation for slides or paper-quality figures, choosing between a large variety of charts, styles and colors. Users can then implement their own algorithms, taking advantage of the high programming capabilities of Visual Basic.

ACKNOWLEDGMENTS

I would like to thank Drs. Jean-Jacques Lebrun and Suad Ali (McGill University, Montréal) for software testing. I thank Prof. J. Gouranton and Dr. D. Thomas (Université de Rennes 1) for their support and for stimulating discussions during the development of this work.

REFERENCES

- Agre, P., M. Bonhivers and M.J. Borgnia. 1998. The aquaporins, blueprints for cellular plumbing systems. *J. Biol. Chem.* 273:14659-14662.
- Aloy, P., J. Cedano, B. Oliva, F.X. Aviles and E. Querol. 1997. 'TransMem': a neural network implemented in Excel spreadsheets for predicting transmembrane domains of proteins. *Comput. Appl. Biosci.* 13:231-234.
- Altschul, S.F., R.J. Carroll and D.J. Lipman. 1989. Weights for data related by a tree. *J. Mol. Biol.* 207:647-653.
- Bairoch, A., P. Bucher and K. Hofmann. 1996. The PROSITE database, its status in 1995. *Nucleic Acids Res.* 24:189-196.
- Briffeuil, P., G. Baudoux, C. Lambert, X. De Bolle, C. Vinals, E. Feytmans and E. Depiereux. 1998. Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance reliability of predictions. *Bioinformatics* 14:357-366.
- Clotet, J., J. Cedano and E. Querol. 1994. An Excel spreadsheet computer program combining algorithms for prediction of protein structural characteristics. *Comput. Appl. Biosci.* 10:495-500.
- Dayhoff, M.O., R.M. Schwartz and B.C. Orcutt. 1978. A model of evolutionary changes in proteins, p. 345-358. *In* M.O. Dayhoff (Ed.), *Atlas of Protein Sequence and Structure*; vol. 5, suppl. 3. National Biochemical Research Foundation, Washington, DC.
- Devereux, J., P. Haeberli and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 12:387-395.
- Eisenberg, D. and A.D. McLachlan. 1986. Solvation energy in protein folding and binding. *Nature* 319:199-203.
- Froger, A., B. Tallur, D. Thomas and C. Delamarche. 1998. Prediction of functional residues in water channel and related proteins. *Protein Sci.* 7:1458-1468.
- Gotoh, O. 1999. Multiple sequence alignment: algorithms and applications. *Adv. Biophys.* 36:159-206.
- Gribskov, M. and R.R. Burgess. 1986. Sigma factors from *E. coli*, *B. subtilis*, phage SP01, and phage T4 are homologous proteins. *Nucleic Acids Res.* 14:6745-6763.
- Han, B. and A.H. Tashjian. 1998. User-friendly and versatile software for analysis of protein hydrophobicity. *BioTechniques* 25:256-263.
- Haygood, M.G. 1993. Spreadsheet macros for coloring sequence alignments. *BioTechniques* 15:1084-1088.
- Henikoff, S. and J.G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89:10915-10919.
- Henikoff, S. and J.G. Henikoff. 1994. Position-based sequence weights. *J. Mol. Biol.* 243:574-578.
- Henikoff, S., J.G. Henikoff, W.J. Alford and S. Pietrokovski. 1995. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* 163:GC17-GC26.
- Hernandez, A. and M.T. Ruiz. 1998. An EXCEL template for calculation of enzyme kinetic parameters by non-linear regression. *Bioinformatics* 14:227-228.
- Heymann, J.B., P. Agre and A. Engel. 1998. Progress on the structure and function of Aquaporin 1. *J. Struct. Biol.* 121:191-206.
- Hofmann, K. 1998. Protein classification and functional assignment, p. 18-21. *Trends Guide to Bioinformatics*. Elsevier Science, New York.
- Ishibashi, K. and S. Sasaki. 1998. The dichotomy of MIP family suggests two separate origins of water channels. *News Physiol. Sci.* 13:137-142.
- Kyte, J. and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105-132.
- Lagré, V., A. Froger, S. Deschamps, I. Pellerin, C. Delamarche, G. Bonnet, J. Gouranton, D. Thomas and J.F. Hubert. 1998. Oligomerization state of water channels and glycerol facilitators. *J. Biol. Chem.* 273:33949-33953.
- Lagré, V., A. Froger, S. Deschamps, J.F. Hubert, C. Delamarche, G. Bonnet, D. Thomas, J. Gouranton and I. Pellerin. 1999. Switch from an aquaporin to glycerol channel by two amino acids. *J. Biol. Chem.* 274:6817-6819.
- Persson, B. and P. Argos. 1994. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.* 237:182-192.
- Rao, J.K.M. and P. Argos. 1986. A conformational preference parameter to predict helices in integral membrane proteins. *Biochim. Biophys. Acta* 869:197-214.
- Schneider, T.D. and R.M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18:6097-6100.
- Schneider, T.D., G.D. Stormo, L. Gold and A. Ehrenfeucht. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188:415-431.
- Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Tech. J.* 27:379-423.
- Shaw, G. 1995. Protein sequence interpretation using a spreadsheet program. *BioTechniques* 19:978-983.
- Thompson, J.D., D.G. Higgins and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.

Received 24 September 1999; accepted 22 December 1999.

Address correspondence to:

Dr. Christian Delamarche
CNRS UPRES-A 6026
Université de Rennes 1
Campus de Beaulieu
Bâtiment 13
35042 Rennes cedex, France
e-mail: cdelam@univ-rennes1.fr