TEXshade: shading and labeling of multiple sequence alignments using LATEX 2ε

Eric Beitz

Faculty of Chemistry and Pharmacy, University of Tübingen, Morgenstelle 8, 72076 Tübingen, Germany

Received on July 8, 1999; revised on September 17, 1999; accepted on September 22, 1999

Abstract

Motivation: *Typesetting, shading and labeling of nucleotide and peptide alignments using standard word processing or graphics software is time consuming. Available automatic sequence shading programs usually do not allow manual application of additional shadings or labels. Hence, a flexible alignment shading package was designed for both calculated and manual shading, using the macro language of the scientific typesetting software LATEX 2ε.*

Results: *TEXshade is the first TEX-based alignment shading software featuring, in addition to standard identity and similarity shading, special modes for the display of functional aspects such as charge, hydropathy or solvent accessibility. A plenitude of commands for manual shading, graphical labels, re-arrangements of the sequence order, numbering, legends etc. is implemented. Further, TEXshade allows the inclusion and display of secondary structure predictions in the DSSP-, STRIDE- and PHD-format.*

Availability: *From http://homepages.uni-tuebingen.de/beitz/tse.html (macro package and on-line documentation)*

Contact: *eric.beitz@uni-tuebingen.de*

Introduction

Gene and protein families are growing rapidly, calling for computer based data management and analysis. The aquaporin superfamily for example, whose first member was identified as a water channel by Preston *et al.* (1992), consists of more than 100 members today (Park and Saier, 1996). Multiple sequence alignments have revealed positions which are responsible for the oligomerization state (Lagrée *et al.*, 1998) and the pore specificity (Lagrée *et al.*, 1999). In parallel, predictions of protein structures and properties based on computer calculations or knowledge databases have become more and more reliable.

In this area, many computational services are provided free on the internet, e.g. multiple sequence alignments with ClustalW (Thompson *et al.*, 1994), secondary structure prediction with PHD (Rost, 1996) or STRIDE (Frishman and Argos, 1995) at http://www.embl-heidelberg.de or at http://helix.nih.gov. As an output the user usually obtains ASCII files. These files have the great advantage of being system independent; nevertheless, they are substantially difficult to survey. Further, an extensive sequence analysis which applies several algorithms results in a multitude of non-connected files. In order to simplify the analysis or to prepare an intelligible presentation these data have to be combined, excerpted, labeled and annotated in a clearly arranged figure.

Manipulating sequence alignments using standard word processing or graphics software is time consuming. Even simple layout changes such as re-breaking lines elongate the working time considerably. Several specialized software tools can format and shade multiple sequence alignments, e.g. Alscript (Barton, 1993), Boxshade (www.isrec.isb-sib.ch/sib-isrec/boxshade), ShadyBox (http://www.angis.su.oz.au/%7echuynh/ShadyBox.html) or ESPript (Gouet *et al.*, 1999). However, these programs fulfill the requirements of cross platform availability, ease of use and flexibility only partially. The highest degree of convenience would be offered by an alignment shading system which is fully integrated into the software routinely used for typesetting scientific texts.

The state-of-the-art scientific typesetting software LATEX 2ε is available for all computer platforms. It has a huge macro language which makes it flexible in a way such that it can be used for setting any kind of text to high quality. Here, a comprehensive LATEX macro package providing more than 100 commands is described. TEXshade has been designed for typesetting, shading and labeling preprocessed multiple sequence alignments in the MSF-file (GCG PileUp) and ALN-file (Clustal) format with the possibility of including secondary structure information from DSSP (Kabsch and Sander, 1983), STRIDE and PHD files. In addition to common shading algorithms for emphasizing identical and similar residues TEXshade provides special shading modes featuring functional aspects, e.g. charge, hydropathy or solvent accessibility and further allows defining new shading modes. In order to offer highest flexibility, a plenitude of commands for implementing manual shading and for handling colours, text styles, labels and legends is available.

System and methods

TeXshade is a macro package which is completely written in TeX/LaTeX 2ε. It has been developed with OzTeX on the Apple Macintosh and has further been tested with emTeX on a Windows 95 Intel-PC. Due to the system independent TeXsource it should run with any recent LaTeX 2ε installation on any system. TeXshade makes intensive use of PostScript for drawing shaded boxes and coloured text using the COLOR.STY package by David Carlisle with the DVIPS option. This package is part of the *Standard LaTeX Graphics Bundle* which is included in every comprehensive LaTeX 2ε distribution. In order to view and print PostScript containing device independent (DVI) files generated by LaTeX 2ε a PostScript compatible TeX viewer and printer are required. Alternatively, the DVI-files can be converted to PostScript using DVIPS by Tomas Rokicki and viewed/printed with a software PostScript interpreter such as GhostView from the GNU free software foundation. All necessary files are available via anonymous ftp from any *Comprehensive TeX Archive Network* (CTAN) server, e.g. ftp.dante.de.

A

```
\begin{texshade}[<parameterfile>]{<alignmentfile>}

    further TeXshade commands, if needed

\end{texshade}
```

B

```
\begin{texshade}{AQPpro.MSF}
    \shadingmode[allmatchspecial]{similar}
    \defconsensus{-}{lower}{upper}
    \showconsensus{top}
    \orderseqs{1,2,4,5,3}
    \separationline{5}
    \setfont{names}{sf}{md}{up}{footnotesize}
    \nameseq{3}{AQP3 (glycerol pore)}
    \namecolor{3}{Gray}
    \numbercolor{3}{Gray}
\end{texshade}
```

C

```
\begin{texshade}{AQPpro.MSF}
    \shadingmode[structure]{functional}
    \feature{top}{1}{185..205}{,->}{extracellular loop E}
    \feature{bottom}{1}{172..184}{--'}{transmembr. Domain 5}
    \feature{bottom}{1}{192..194}{brace}{NPA}
    \setfont{features}{sf}{md}{up}{small}
\end{texshade}
```

Fig. 1. Usage of the TeXshade environment in a LaTeX 2ε document. A. A minimum environment definition contains no further TeXshade commands and uses the default settings (for an output see Figure 2A). B. and C. Additional commands which were used for the examples in Figures 2B and 2E.

Algorithm and implementation

The TeXshade environment

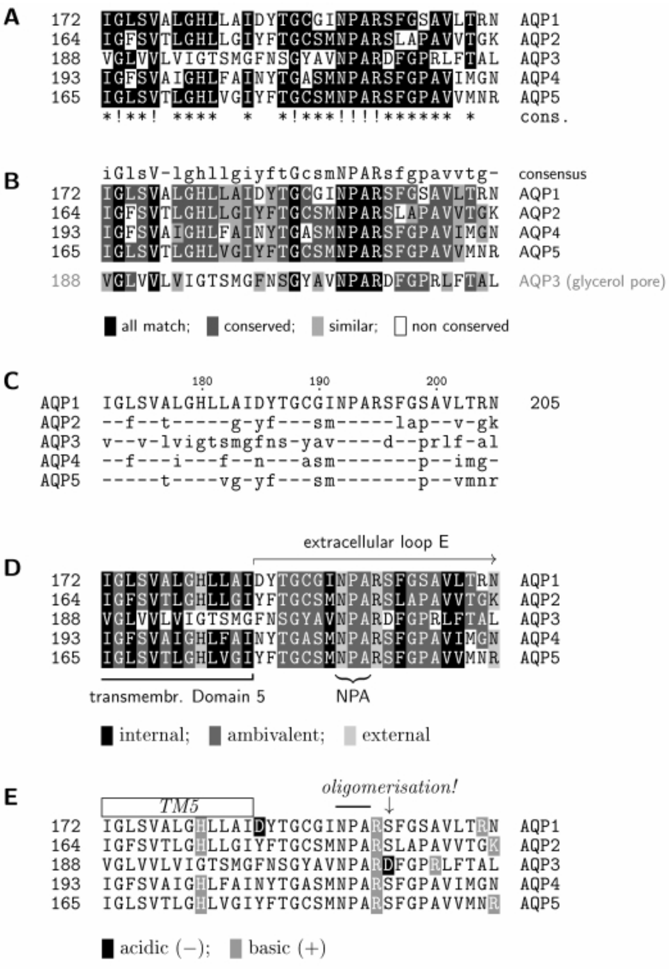
A new LaTeX 2ε environment for setting alignments is provided by TeXshade whose usage is similar to well known standard LaTeX environments like ‘tabular’ or ‘itemize’. At the beginning of the environment the filename of a preprocessed alignment file in the MSF or ALN format is designated. An optional argument takes the filename of a user specific parameter file which can contain any TeXshade command. Parameter files are intended for the management of standard settings and user defined shading modes. Within the environment further TeXshade commands determine the final appearance of the alignment. Figure 1 shows the usage of the TeXshade environment.

Identity, similarity and diversity shading

Sequences can either be compared to a single master sequence or a consensus sequence is calculated from the residue fraction at a position which exceeds a given threshold. In the most basic mode, the so-called *identity mode*, only identical residues at a position are shaded. More information about sequence relationships can be obtained in the *similarity mode*. Here, additional residues are shaded which are (a) similar to the consensus residue or (b) exceed the threshold as a group of similar residues. The default definitions for similarity were adopted from Boxshade. A user specific customization is possible. For both identity and similarity mode a special shading colour can be displayed at positions where all residues are identical. The *diversity mode* is implemented for the demonstration of differences between closely related sequences, e.g. species variants of a protein. Here, residues which are identical with a master sequence are blanked out, whereas non-matching residues are displayed. Shading examples for the above described modes are shown in Figures 2A–C.

Functional shading modes

Six *functional modes* are predefined in the TeXshade macro package. In contrast to identity or similarity comparisons there is no calculation of a single consensus residue but of a consensus group of residues which are functionally similar. The members of the consensus group are labeled with the respective group colour, see Figure 2D. An overview about the amino acid grouping is given in Table 1. *Functional mode* allows ignoring the consensus calculation resulting in the shading of all residues, whether matching, or non-matching, due to their group assignment. For instance, with this option activated all charged residues or information about the solvent accessibility throughout all sequence positions can be displayed (Figure 2E). TeXshade provides the necessary



```
A
172  IGLSVALGHLLAIDYTGCGINPARSFGSAVLTRN  AQP1
164  IGFSVTLGHLLGIYFTGCSMNPARSLAPAVVTGK  AQP2
188  VGLVVLVIGTSMGFNSGYAVNPARDFGPRLFTAL  AQP3
193  IGFSVAIGHLFAINYTGASMNPARSFGPAVIMGN  AQP4
165  IGLSVTLGHLVGIYFTGCSMNPARSFGPAVVMNR  AQP5
     *!**! ****  *  *!***!!!!****** *    cons.

B
     iGlsV-lghllgiyftGcsmNPARsfgpavvtg-  consensus
172  IGLSVALGHLLAIDYTGCGINPARSFGSAVLTRN  AQP1
164  IGFSVTLGHLLGIYFTGCSMNPARSLAPAVVTGK  AQP2
193  IGFSVAIGHLFAINYTGASMNPARSFGPAVIMGN  AQP4
165  IGLSVTLGHLVGIYFTGCSMNPARSFGPAVVMNR  AQP5
188  VGLVVLVIGTSMGFNSGYAVNPARDFGPRLFTAL  AQP3 (glycerol pore)

C
               180       190       200
AQP1  IGLSVALGHLLAIDYTGCGINPARSFGSAVLTRN   205
AQP2  --f--t-----g-yf---sm-----lap--v-gk
AQP3  v--v-lvigtsmgfns-yav----d--prlf-al
AQP4  --f---i---f--n---asm-------p--img-
AQP5  -----t----vg-yf---sm-------p--vmnr

D
172  IGLSVALGHLLAIDYTGCGINPARSFGSAVLTRN  AQP1
164  IGFSVTLGHLLGIYFTGCSMNPARSLAPAVVTGK  AQP2
188  VGLVVLVIGTSMGFNSGYAVNPARDFGPRLFTAL  AQP3
193  IGFSVAIGHLFAINYTGASMNPARSFGPAVIMGN  AQP4
165  IGLSVTLGHLVGIYFTGCSMNPARSFGPAVVMNR  AQP5

E
172  IGLSVALGHLLAIDYTGCGINPARSFGSAVLTRN  AQP1
164  IGFSVTLGHLLGIYFTGCSMNPARSLAPAVVTGK  AQP2
188  VGLVVLVIGTSMGFNSGYAVNPARDFGPRLFTAL  AQP3
193  IGFSVAIGHLFAINYTGASMNPARSFGPAVIMGN  AQP4
165  IGLSVTLGHLVGIYFTGCSMNPARSFGPAVVMNR  AQP5
```

■ all match; ■ conserved; ■ similar; □ non conserved
extracellular loop E
transmembr. Domain 5
NPA
■ internal; ■ ambivalent; ■ external
TM5
oligomerisation!
■ acidic (−); ■ basic (+)

Fig. 2. Example outputs of different shading modes. A highly conserved region of aquaporin types 1–5 from rat is shown, including the asparagine/proline/alanine motif which is part of the pore region. A. *Identity mode*, B. *similarity mode*, C. *diversity mode*, D. *functional mode*: structure, E. *functional mode*: charge with consensus ignored. All alignments were set using the same alignment input file. The different kinds of shading and labeling are produced quickly by only a few commands. Note the sequence order in B and further layout changes. Rearrangements of the sequences and the display of alignment sections are possible without editing the input file.

commands for the definition of additional functional groups with any colour and describing text for the display in the legend.

Including secondary structure information

Secondary structure calculations in the DSSP, STRIDE and PHD format can be included into the alignment output by a single command. Also, PHD files provide topology predictions of membrane proteins indicating intracellular, transmembrane and extracellular regions. TEXshade extracts all positions of helices, strands and turns from the prediction file and converts the information into TEXshade commands. These are stored in a new file which can be fully edited for customization. The number of files to be included is not limited, i.e. a

Table 1. Amino acid grouping due to functional aspects predefined in TEXshade

Mode	Group	Residues
`charge`[a]	acidic (−)	D, E
	basic (+)	H, K, R
`chemical`[a]	acidic (−)	D, E
	aliphatic	A, G, I, L, V
	amide	N, Q
	aromatic	F, W, Y
	basic (+)	H, K, R
	hydroxy	S, T
	imino	P
	sulfur	C, M
`structure`[a]	ambivalent	A, C, G, P, S, T, W, Y
	externa	D, E, H, K, N, Q, R
	internal	F, I, L, M, V
`hydropathy`[b]	acidic (−)	D, E
	basic (+)	H, K, R
	hydrophobic	A, F, I, L, M, P, V, W
	polar uncharged	C, G, N, Q, S, T, Y
`standard area`[c]	$< 110\,\text{Å}^2$	G
	$110–129\,\text{Å}^2$	A, S
	$130–149\,\text{Å}^2$	C, P
	$150–169\,\text{Å}^2$	D, N, T, V
	$170–189\,\text{Å}^2$	E, I
	$190–209\,\text{Å}^2$	H, L, M, Q
	$210–229\,\text{Å}^2$	F, K
	$230–249\,\text{Å}^2$	Y
	$> 249\,\text{Å}^2$	R, W
`accessible area`[c]	$< 18\,\text{Å}^2$	C
	$18–27\,\text{Å}^2$	G, I, V
	$28–37\,\text{Å}^2$	A, F, L, M
	$38–47\,\text{Å}^2$	H, S, T, W
	$48–57\,\text{Å}^2$	P
	$58–67\,\text{Å}^2$	D, N, Y
	$68–77\,\text{Å}^2$	E, Q
	$78–97\,\text{Å}^2$	R
	$> 97\,\text{Å}^2$	K

Each group of a shading mode is labeled with a different colour. The residue grouping of the modes ‘charge’, ‘chemical’ and ‘hydropathy’ is self-explanatory. ‘Structure’ differentiates between residues which are preferentially localized in the core of a globular protein and those presented on the surface. ‘Standard area’ visualizes the different side chain areas of the residues, whereas ‘accessible area’ displays the statistically evaluated solvent accessibility of each amino acid in a folded protein.
[a] Karlin and Ghandour, 1985.
[b] Kyte and Doolittle, 1982.
[c] Rose *et al.*, 1985; Lesser and Rose, 1990.

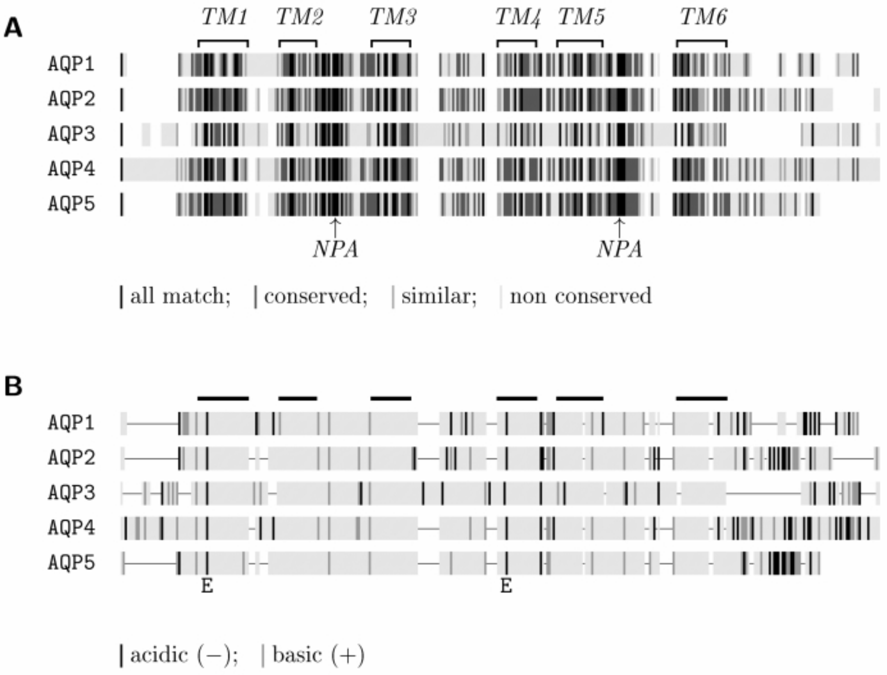


Fig. 3. Display of two sequence fingerprints of aquaporins 1–5. A. The most conserved stretches are found in the transmembrane regions and around the NPA-motifs indicated by the dark shading (*similarity mode*). B. The accumulation of charged amino acids is highest near the C terminus, whereas the major parts of the proteins are uncharged due to the transmembrane topology (*functional mode*: charge). Note the two conserved glutamic acid residues in transmembrane domains 1 and 4.

direct comparison of different prediction algorithms in the same alignment is possible. Labels for secondary structures can be chosen from sets of arrows, bars, braces, boxes and so-called fill characters. The latter are symbols or letters which are printed serially until the respective sequence stretch is fully covered. Figures 2D–E show some examples. Four individual rows, two at the top and two at the bottom of the alignment, are reserved for the display of labels. Thus, up to four labels for one sequence position or region can be shown at a time.

Fingerprinting

An easy way to gain overviews on complete alignments is provided by the display of an alignment *fingerprint*. In this mode the whole sequence can be shown in one line. Due to the lack of space the residue symbols are hidden and the shaded boxes are reduced to thin vertical lines. The result is a horizontal bar with coloured vertical lines as a representation of the sequence (Figure 3). All TEXshade commands are compatible with fingerprinting, i.e. all shading modes are applicable for displaying overviews on sequence similarity or functional aspects. Also, the inclusion of secondary structure information and all kinds of labeling is possible in conjunction with this command.

Manual labeling and further commands

In addition to computer calculated shading and labeling there is the possibility of manually setting graphic labels such as those described above. Further, positions or regions within individual sequences can be emphasized by applying distinct colours and text styles. The positions of the labels are assigned by the residue numbers within the sequence. This means, that the label is tightly attached to the sequence stretch and moves with the sequence whenever the line layout changes, e.g. by changing the number of residues per line.

In general, the colour and text style of any text in the alignment (sequence names, numbering, feature descriptions, legend and residues) is freely selectable. The standard LATEX font families plus the full set of PostScript fonts is applicable.

TEXshade allows the rearrangement of the sequence order and the display of sections of the preprocessed input file without the need to edit or to recalculate the entire alignment. Sequences can also be hidden in the output with or without consideration for the calculation of the consensus. Vertical gaps can be inserted in order to indicate subfamilies (see Figure 2B) and the numbering can be manipulated if the sequence does not start out at position 1.

Discussion

System independency, ease of use and flexibility, combined with high output quality, were the major aims for the development of the TEXshade package. Two aspects, system independency and output quality, were achieved by the choice of the public domain typesetting software LATEX 2$_\varepsilon$ which is available for all computer platforms. TEXshade is the first molecular biology extension for LATEX 2$_\varepsilon$ and thus opens a new field of application to TEX. Setting alignments with TEXshade is easy, because the command syntax corresponds to the TEX conventions and due to comprehensive default settings the standard output is easily obtained by an empty environment definition as shown in Figure 1. This output is step-by-step fitted to the user's ideas by including more commands into the environment. The nine predefined shading modes are helpful, covering the major needs without a great deal of typesetting.

A personal parameter file which is loaded and executed at the beginning of the TEXshade environment enhances the efficiency of the macro package. Whenever a satisfactory shading has been created by the user, the corresponding commands can be stored as a parameter file for later use or for exchange with others. Using personal standard settings throughout a publication or presentation leads to an individual, but consistent, overall appearance.

The option to include secondary structure information from DSSP, STRIDE or PHD files by a single command is convenient. Thus, there is no need to apply manual labels for the indication of secondary structures. To provide enough flexibility, the conversion into TEXshade commands results in a file similar to a parameter file which is fully editable for individual customization.

Fingerprints calculated by TEXshade are similar to the output obtained from the Macintosh HyperCard software *Sequence Similarity Presenter* by Fröhlich (1994). With this software the alignment is represented as horizontal bars filled with shades of gray of varying intensity due to the degree of similarity. The smooth edged shading is achieved by the calculation of mean values from a moving window covering several vicinal residues. TEXshade in contrast compares residue per residue leading to a less smooth but 'bar code like' output. Nevertheless, it provides the possibility of not only displaying sequence similarities but also functional shading.

The user has maximal influence on parameters determining the calculation of the shading and the appearance of manual labels. The macro package is open for almost any change, especially the definition of new shading modes; by re-grouping the residues the macro can be adapted to many tasks. Further, where computational shading is not sufficient, additional manual shading can be applied easily. This, taken together with other possibilities like re-arrangement of the sequence order, sectioning of the alignment or manipulation of the numbering makes TEXshade a highly flexible tool.

In conclusion, TEXshade is an alignment shading software suitable for both sequence analysis and presentation.

References

Barton,G.J. (1993) ALSCRIPT, A tool to format multiple sequence alignments. *Protein Eng.*, **6**, 37–40.

Frishman,D. and Argos,P. (1995) Knowledge-based secondary structure assignment. *Proteins*, **23**, 566–579.

Fröhlich,K.-U. (1994) Sequence Similarity Presenter: a tool for the graphic display of similarities of long sequences for use in presentations. *Comput. Appl. Biosci.*, **10**, 179–183.

Gouet,P., Courcelle,E., Stuart,D.I. and Métoz,F. (1999) ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics*, **15**, 305–308.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Karlin,S. and Ghandour,G. (1985) Multiple-alphabet amino acid sequence comparisons of the immunoglobulin κ-chain constant domain. *Proc. Natl Acad. Sci. USA*, **82**, 8597–8601.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of protein. *J. Mol. Biol.*, **157**, 105–132.

Lagrée,V., Froger,A., Deschamps,S., Hubert,J.-F., Delamarche,C., Bonnec,G., Gouranton,J. and Pellerin,I. (1999) Switch from an aquaporin to a glycerol channel by two amino acids substitution. *J. Biol. Chem.*, **274**, 6817–6819.

Lagrée,V., Froger,A., Deschamps,S., Pellerin,I., Delamarche,C., Bonnec,G., Gouranton,J., Thomas,D. and Hubert,J.-F. (1998) Oligomerization state of water channels and glycerol facilitators. *J. Biol. Chem.*, **273**, 33949–33953.

Lesser,G.J. and Rose,G.D. (1990) Hydrophobicity of amino acid subgroups in proteins. *Proteins*, **8**, 6–13.

Park,J.H. and Saier,Jr.,M.H. (1996) Phylogenetic characterization of the MIP family of transmembrane channel proteins. *J. Membrane Biol.*, **153**, 171–180.

Preston,G.M., Carroll,T.P., Guggino,W.B. and Agre,P. (1992) Appearance of water channels in *Xenopus* oocytes expressing red cell CHIP28 protein. *Science*, **256**, 385–387.

Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985) Hydrophobicity of amino acid residues in globular proteins. *Science*, **229**, 835–838.

Rost,B. (1996) PHD: prediction one-dimensional protein structure by profile based neural networks. *Methods Enzymol.*, **266**, 525–539.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W.: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, **22**, 4673–80.