Biochimica et Biophysica Acta 1818 (2012) 907–914

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbamem

Identification of residues involved in water versus glycerol selectivity in aquaporins by differential residue pair co-evolution

Xin Lin [a,b,c], Tian Hong [a], Yuguang Mu [a], Jaume Torres [a,*]

[a] School of Biological Sciences, Nanyang Technological University, Singapore
[b] School of Civil and Environmental Engineering, Nanyang Technological University, Singapore
[c] Singapore Membrane Technology Centre, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 24 June 2011
Received in revised form 15 December 2011
Accepted 20 December 2011
Available online 29 December 2011

Keywords:
Water channel
Aquaporin selectivity
Co-evolution analysis
Binary classifier
Water versus glycerol selectivity
Gating of aquaporins

ABSTRACT

Aquaporins (AQPs) are members of the Major Intrinsic Protein (MIP) family that can transport water or glycerol, as well as other compounds. The rationale for substrate selectivity at the structural level is still incompletely understood. The information present in multiple sequence alignments (MSAs) can help identify both structural and functional features, especially the complex networks of interactions responsible for water or glycerol selectivity. Herein, we have used the method of Statistical Coupling Analysis (SCA) to identify co-evolving pairs of residues in two separate groups of sequences predicted to correspond to water or glycerol transporters. Differentially co-evolved pairs between the two groups were tested by their efficacy in correctly classifying a training set of MSAs, and binary classifiers were built with these pairs. Up to 50% of the residues found in hundreds of binary classifiers corresponded to only ten positions in the MSA of aquaporins. Most of these residues are close to the lining of the aquaporin pore and have been identified previously as important for selectivity. Therefore, this method can shed light on the residues that are important for substrate selectivity of aquaporins and other proteins. SCA requires a very large sequence dataset with relatively low homology amongst its members, and these requirements are met by aquaporins.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Aquaporins are a family of small (28–30 kDa) pore-forming integral membrane proteins and are important actors of fluid homeostasis [1]. The ancient name of this family (major intrinsic proteins, MIP) derives from a protein, MIP26, found in mammalian lens fibers [2,3] — now designated AQP-0. MIP homologs were later shown to function as water channels, hence the name 'aquaporins'. Generally, MIP homologs with exclusive water permeability are referred to as aquaporins (AQPs), whereas those permeable to both water and glycerol are referred to as glycerol facilitator proteins (GLP). In all aquaporins, transport is a passive mechanism driven by the concentration gradient. Aquaporins are found in all organisms, from bacteria to humans [4–10], although with different distribution: whereas many eubacteria have a single AQP and a single GLP, there are 13 aquaporins in humans (AQP0–AQP12). Of these, 0, 1, 2, 4, 5, 6, 8 are water channels, 3, 7, 9, 10 are aquaglyceroporins, whereas 11 and 12 are termed 'superaquaporins' (reviewed in [11]). However, the separation between these categories is blurred due to the multiplicity of substrates used by a given aquaporin. Indeed, aquaporins can permeate nitrate and chloride ions [12,13], ammonia [14,15], glycerol and urea [16] or toxic metalloids such as arsenite and antimonite [17,18]. It has also been suggested that gases like CO_2, O_2 and NO can be also permeated by aquaporins [19,20]. In plants, MIP genes are particularly abundant; more than 35 different genes encoding aquaporin-like proteins are found in *Arabidopsis thaliana* [21,22].

Sequence identity is low in aquaporins; for example, between the human water channel protein AQP1 and the bacterial aquaglyceroporin GlpF, identity is less than 30.6% [23]. The first aquaporin member described was the 28 kDa protein of the human erythrocyte membrane, later renamed as aquaporin-1 (AQP1) [24]. Since then, structures of several MIPs have accumulated in the Protein Data Bank, e.g., mammalian aquaporins AQP0 [25], AQP1 [26], AQP4 [27] and AQP5 [28], bacterial aquaporins GlpF [29], AqpM [30], AqpZ [31], and plant aquaporins, e.g., SoPIP2 [32]. Aquaporins share a similar general structure, with six transmembrane (TM) domains connected by five loops (A–E), where both N- and C-termini are intracellular. The structure has two similar halves that probably arose by gene duplication [33]: TMs 1–3 form the so-called hemipore-1, and TMs 4–6 form hemipore-2. These two hemipores face each other inside the membrane, forming an hourglass-like shape. Loops B and E form short hydrophobic helices that penetrate into the membrane from opposite sides, and contain highly conserved NPA (Asn-Pro-Ala) motifs that are located in the center of the monomer pore and may participate in substrate selectivity [34]. MIPs are

* Corresponding author. Tel.: +65 6316 2857; fax: +65 6791 3856.
E-mail address: jtorres@ntu.edu.sg (J. Torres).

0005-2736/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbamem.2011.12.017

arranged as homotetramers, and each monomer functions independently as a pore.

The elucidation of the structure of aquaporins [35] and aquaglyceroporins [36] gave the first insights into their selectivity mechanisms [29]. Differences in channel selectivity [37] are determined by charge, polarity and size [38]. Mainly, there are two constriction points within the pore: the Asn-Pro-Ala (NPA) motif and the aromatic/arginine (ar/R) selectivity filter, which impairs the entry of high molecular weight substrates, and constitutes a checkpoint for uncharged molecules in both AQPs and aquaglyceroporins. The amino acids in and around that filter may provide hydrogen bonds that confer high selectivity for water transport [39,40] and can also influence the polarity and the diameter of the pore. The divergence of these amino acids among MIP isoforms is thought to constitute the major difference between AQPs and aquaglyceroporins.

Multiple sequence alignments (MSA) of proteins constitute a rich source of information, such as residue conservation or hydrophobicity. Froger et al. [41] attempted to find key functional residues that separate the two groups of aquaporins on the basis of differences in physico-chemical characteristics at certain positions in MSAs. However, reversal of substrate specificity by point mutagenesis has only resulted in partial success [37,42]. This suggests that selectivity is likely to reside in a complex network of interactions of residues, where some may not be in contact with the substrate.

More complex relationships in MSAs are detected when a residue co-evolution analysis is performed. The latter detects two or more positions in the MSA that may not be overall conserved, yet experience a synchronous change in composition, being indicative of functional or structural importance. Herein lays the main advantage of co-evolution methods, which has applications in prediction of folding, interacting domain between two proteins, or binding/functional sites, where networks of correlated mutations appear.

Statistical Coupling Analysis (SCA) [43] is one of the methods available to discover co-evolving residues, and requires a large and varied dataset, i.e., hundreds of sequences with low conservation. As the number of MIP sequences available is very large, with up to 2035 proteins in 2009 in the Pfam database (http://pfam.sanger.ac.uk) [44], it is therefore adequate for this analysis. In this method, statistical 'coupling' between two sites, i and j, (represented by two columns in the MSA) is detected if they exert mutual evolutionary pressure, which leads to a distribution of amino acids at positions i and j that deviates from the unconstrained distribution found for the whole MSA. Mutual dependence is measured by performing a 'perturbation experiment', where a subset of sequences in the MSA containing a certain amino acid at position i is selected. For this subset, if a coupling exists between sites *i* and *j*, a bias in the amino acid distribution at site j should be observed. The magnitude of this bias is quantitatively measured as a statistical coupling energy between these two sites [43,45,46].

As in Froger et al. [41], we attempt to find key functional residues that separate the two groups of aquaporins, water or glycerol transporters, but we do that by detecting co-evolving pairs of residues in either group, rather than by side by side comparison. As outlined above, we formed two groups of aquaporin sequences, classified previously with phylogenetic methods [34,47], and assigned to 'water channels' (AQP) or 'glycerol transporters' (GLP). For each class of sequences, we used SCA to search for coupled pairs, and those pairs that showed significantly different degree of coupling between the two groups, i.e., coupled in one group and uncoupled in the other, were thought to be likely important for function. The ability of the residues in those pairs to correctly classify aquaporin sequences in either of these two groups was used to test the relevance of those residues in representing that particular group, regardless of their functional meaning. For example, if a coupled pair was often found in the AQP group, but not in the GLP group, comparison of the amino acids in that pair with the amino acids in the interrogated sequence should lead to assignment of that sequence to the AQP group. We selected more relevant pairs from the pool by adding those that increased the accuracy of the classification. Our hypothesis was that the positions in the MSA represented by the coupled pairs of residues, which form a 'binary classifier', must represent key residue positions that determine substrate specificity. Comparison of these identified key residues with experimentally confirmed key sites related to water/glycerol selectivity and gating confirms our hypothesis.

2. Material and methods

2.1. Multiple sequence alignment (MSA) of aquaporin

The aquaporin sequences for the SCA analysis were obtained using BLAST searches against a non-redundant protein database. First, 14 aquaporin sequences representing unique aquaporin types: AQP0 (NCBI protein ID: NP_036196), AQP1 (NCBI protein ID: NP_932766), AQP2 (NCBI protein ID: CAG46821), AQP3 (NCBI protein ID: CAG46822), AQP4 (NCBI protein ID: NP_001641), AQP5 (NCBI protein ID: CAG46819), AQP6 (NCBI protein ID: NP_001643), AQP7 (NCBI protein ID: NP_001161), AQP8 (NCBI protein ID: NP_001160), AQP9 (NCBI protein ID: CAG46824), AQP10 (NCBI protein ID: CAH70483), AQPM (NCBI protein ID: NP_275246), AQPZ (NCBI protein ID: NP_752939), GLPF (NCBI protein ID: NP_290556), were selected from the database. Subsequently, a PSI-BLAST [48] ($e<0.001$) was run for each of these 14 sequences to generate groups of more than 3000 homologous sequences for each type. Combining the results for these 14 sequences, and after removing identical sequences, a set with 3269 homologous sequences was obtained. Only those sequences that in their description of the database entries had been annotated as either "water transporters" or "glycerol facilitators" were selected, resulting in only 985 sequences. The first class contained 437 sequences, including AQP0, AQP1, AQP2, AQP4, AQP5, AQP6, AQP8 and AQPZ. The second class contained 548 sequences, including AQP3, AQP7, AQP9, AQP10, and GLPF. Multiple sequence alignment (MSA) of these 985 sequences was performed using ClustalW [49]. Sequences with high identity (>90%) were removed from the set, leaving two groups of around 300 sequences each, of either "glycerol facilitator" or "water transporter" sequences. Finally, 219 sequences were randomly selected from each group so that both classes, glycerol facilitator and water transporters, contained the same number of sequences. To increase the significance of the coupling analysis, columns in the MSA (i.e., positions) containing more than 30% of gaps were not used during the calculations, leaving only 192 available columns in the MSA for analysis.

2.2. Statistical coupling analysis

As described previously [50], the coupling between sites *i* and *j* is calculated as a statistical energy $\Delta\Delta G_{i,j}{}^{stat} \approx \frac{1}{f_i^{(a_i)}} \frac{\partial D_j^{(a_j)}}{\partial f_j^{(a_j)}} |C_{ij}|$, where $f_i^{(a_i)}$ is the frequency of amino acid *a* at site *i*, $D_j^{(a_j)}$ is the so-called 'relative entropy', a measure of 'positional conservation' of amino acid *a* at site *j*, and C_{ij} is the reduced weight matrix which represents the positional correlation between sites *i* and *j*. All the calculations were performed using an adapted version of the SCA Toolbox distribution, SCA v3.0 [50].

2.3. Comparison of residue coupling between the two classes

Coupling matrices were obtained from each of the MSA of the two groups of sequences, and were later compared. A schematic representation of the process is shown in Fig. 1, where coupling matrices for water transporters (AQPs, Fig. 1A) and glycerol facilitators (GLPs,

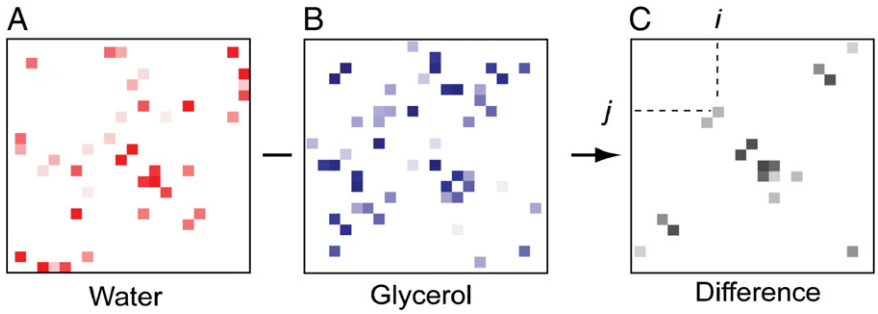

Fig. 1. Schematic representation of the method used to detect differentially co-evolved pairs of residues. (A), SCA matrix corresponding to the water transporter class (AQP); (B), SCA matrix corresponding to the glycerol facilitator class (GLP). High degree of coupling in A and B is shown by intensity of red and blue squares, respectively; (C), Difference between matrices A and B (gray squares), showing the candidate pairs to be used for generating binary classifiers.

Fig. 1B) are compared, and the difference matrix between A and B (Fig. 1C) represents pairs (i, j) with a high degree of coupling in one class (AQPs) and a low degree of coupling in the other (GLPs), or vice versa. Therefore, our hypothesis is that if a pair (i, j) appears in panel C, this pair may be important for substrate selectivity. For every pair of positions (i, j) in the MSA that showed high differential coupling (Fig. 1C), we selected the two most frequently observed pairs of residues, e.g., for positions 30 and 120, we could select the pairs (A30, G120) and (H30, E120).

To confirm the relevance of these pairs of residues, we tested their ability to classify aquaporin sequences belonging to the two groups, and additional pairs were added to that initial pair in order to increase the accuracy of classification until no further improvement in accuracy was achieved. The pairs in that group constituted a 'binary classifier'. We note that although one possible approach would be to just combine these pairs into one large binary classifier, we found that such classifier showed low accuracy (not shown). Therefore, the binary classifier was formed by populating it with pairs one at a time, iteratively.

2.4. Classification of sequences

The classification of an aquaporin sequence with a binary classifier formed by N pairs of coupled residues was based on the S score, $S = \sum_{a=1}^{N} S_a$, where N is the number of pairs in the classifier, and S_a is the score of the sequence using the ath pair of the classifier. The score was calculated as: $S_a = \delta(R_{pair_a}, W_{pair_a}) - \delta(R_{pair_a}, G_{pair_a})$ where R_{pair_a} is the pair of amino acids present at the two positions (i, j) in the aquaporin sequence being classified, whereas W_{pair_a} and G_{pair_a} are pairs of amino acids being tested, from either the water transporter MSA (W_{pair}) or the glycerol transporter MSA (G_{pair}). For example, if the pair (A30, G120) has been found with high abundance in water transporters, one would check the amino acids in the sequence to be classified at equivalent positions 30 and 120 in the MSA. Comparing the pair of residues in that sequence with the probe pair (A30, G120), S_a is assigned $+1$ if the residues in the aquaporin sequence are identical to those of the probe pair, 0 if the pair of residues in the sequence does not match the probe pair, and -1 if they match the probe pair derived from glycerol transporters. Hence, S_a is the subtraction of two Kronecker delta functions. A positive, negative, or 0 value for S determines that the sequence is classified as water transporter, glycerol facilitator, or ambiguous, respectively. The accuracy of the classification was calculated as:

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

where ACC is the accuracy, TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the false negatives.

2.5. Building a classifier using the selected candidate pairs

To build the binary classifiers, a 'training set' of 300 sequences was generated by randomly combining 150 sequences from each class, i.e., water transporters (AQP) and glycerol-facilitators (GLP). At the same time, a 'testing set' was generated by combining 30 sequences taken randomly from either class, using sequences not included in the 'training set', i.e., no overlap existed between 'training set' and 'testing set' sequences. The first pair of residues to be included in the binary classifier was selected by choosing the pair with highest accuracy in the classification of the training sets. From the pool of candidate pairs, a new pair was added iteratively to the binary classifier only if it increased the previously achieved accuracy (see Fig. 2). This was repeated with available candidate pairs until the accuracy of the binary classifier could no longer be increased by addition of more candidate pairs. Each classifier was thus formed by a group of residue pairs. When the above procedure was repeated starting from a different 'training set' of aquaporin sequences, a different binary classifier, i.e., formed by different pairs, usually was obtained. The building procedure was repeated 1000 times to ensure no classifier was missed, which resulted in 718 classifiers with a high degree of overlap between them.

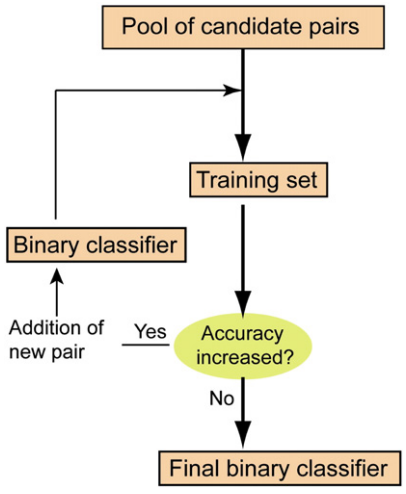

Fig. 2. Flow chart of binary classifier generation. The pool of candidate pairs was tested against a training set, and only pairs that increased accuracy in classification were added to a growing binary classifier. This cycle was restarted 1000 times. Each cycle, a new binary classifier was obtained, which contained a number of coupled pairs (elements).

3. Results and discussion

3.1. Generation of binary classifiers

As described in the Material and Methods section, we selected the coupled pairs that best *represent* the two phylogenetically formed groups of sequences by testing their efficiency in classifying aquaporin sequences into the AQP or the GLP group (Fig. 2). By repeating the training to obtain classifiers 1000 times, a total of 718 different classifiers were obtained, each comprising of 2 to 8 coupled pairs (henceforth referred to as elements), although most classifiers (659 out of 718) contained only 3 to 5 elements (Fig. 3A). The accuracy of these classifiers ranged was 88–93%, which is comparable to classification methods such as decision trees [51], discriminate analysis [52], neural networks and support vector machines, which range from 86% to 97% [53–57]. However, we stress the point that our objective here was neither to classify aquaporin sequences, nor to determine substrate selectivity in those sequences. The goal was to assess the ability of the coupled pairs chosen to represent either group of pre-classified aquaporins. It is for this reason that we have not used sequences that are classified as ambiguous, because they would not provide the desired supervisory signal. We also note that, in this study, the classification is carried out only in a subspace of aquaporin substrate selectivity. This means that there may be many other selectivities (for other substrates) which for simplicity we have not considered. To account for possible wrongly classified sequences, we introduced a percentage of 'wrong' sequences in the training set. We found that the overall accuracy remained at a high level, especially when the percentage of 'wrong' sequences was below 10% (not shown).

Although we found that accuracy increased with the number of elements, accuracy could not increase further when the number of elements was higher than 8 (Fig. 3B). The number of non-identical binary classifiers, i.e., those with at least one non-common element, increased almost linearly when increasing the number of trials (Fig. 3C). However, after a certain number of classifiers had been obtained, new classifiers were simply combinations of previously found ones, therefore they did not contribute to find new co-evolved pairs from the pool. The accuracy in classification of these binary classifiers ultimately depended on the precise combination of the elements forming the classifier and the *order* in which they were tested, i.e., total accuracy was not an additive property of the elements forming the classifier. Therefore, the best binary classifier was not formed necessarily by a combination of the elements that individually produced more accuracy.

3.2. Identification of key residues that determine water versus glycerol permeability

Out of the 192 MSA positions included in the analysis, only 134 (~69%) were represented in at least one of the 718 binary classifiers found. Further, after 1000 trials, the number of positions in the MSA that accounted for 90% and 50% of all the elements found in all the classifiers was 54 and 10, respectively (Fig. 3D). In contrast, the other 50% of the elements found in the classifiers could only be formed by combination of another 132 MSA positions. These numbers were almost unchanged from 100 to 1000 trials (Fig. 4E). The 10 MSA positions referred to above (Table 1) were assumed to have a key role in the determination of substrate permeability and were analyzed in more detail.

3.3. Physico-chemical differences of ten key residues between AQP and GLP sequences

We speculated that these 10 positions would show different amino acid distributions in the groups AQP and GLP (Fig. 4). At some positions, e.g., 26, 102, 144, 160 and 164, the differences in distribution for charged, polar and hydrophobic amino acids is evident. However, the distribution of amino acids was still markedly different at other positions, despite grouping into 'families' results in apparently similar patterns. For example, at position 22, Val, Leu or Ile appeared in ~65% of the sequences in the AQP class, whereas in the

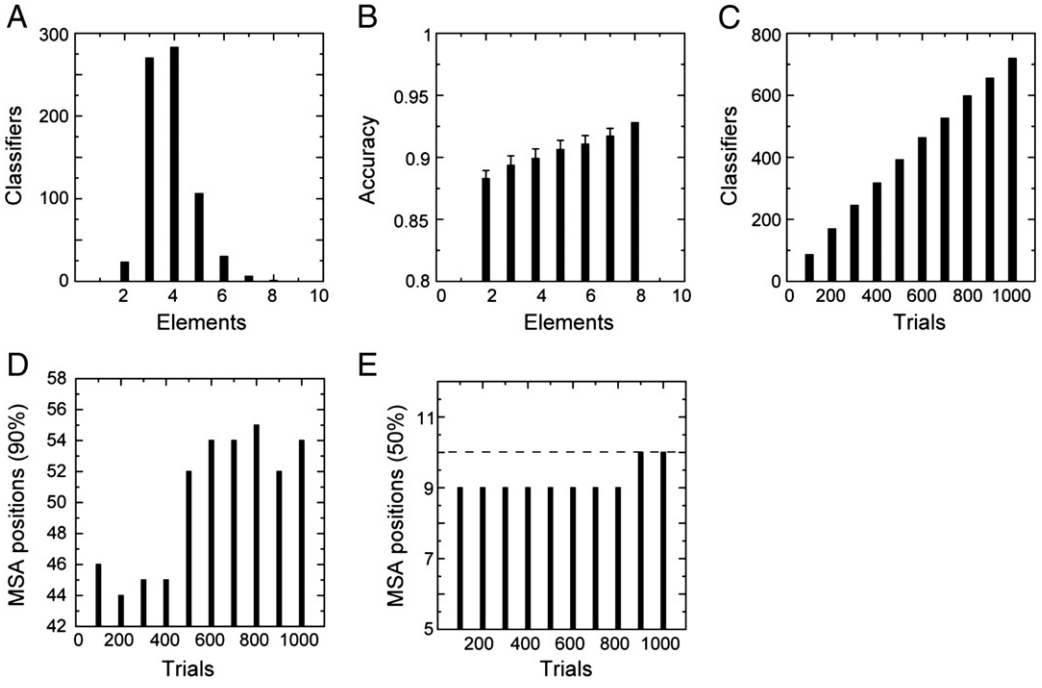

Fig. 3. Statistics of non-identical binary classifiers obtained after 1000 cycles of training. (A), Distribution of binary classifiers according to their number of elements; (B), accuracy of the binary classifiers as a function of number of elements; (C), number of non-identical binary classifiers obtained as a function in number of trials performed; (D–E), number of MSA positions present in the binary classifiers that accounted for either 90% (D), or 50% (E) of the unique positions found amongst all binary classifiers, as a function of number of trials.

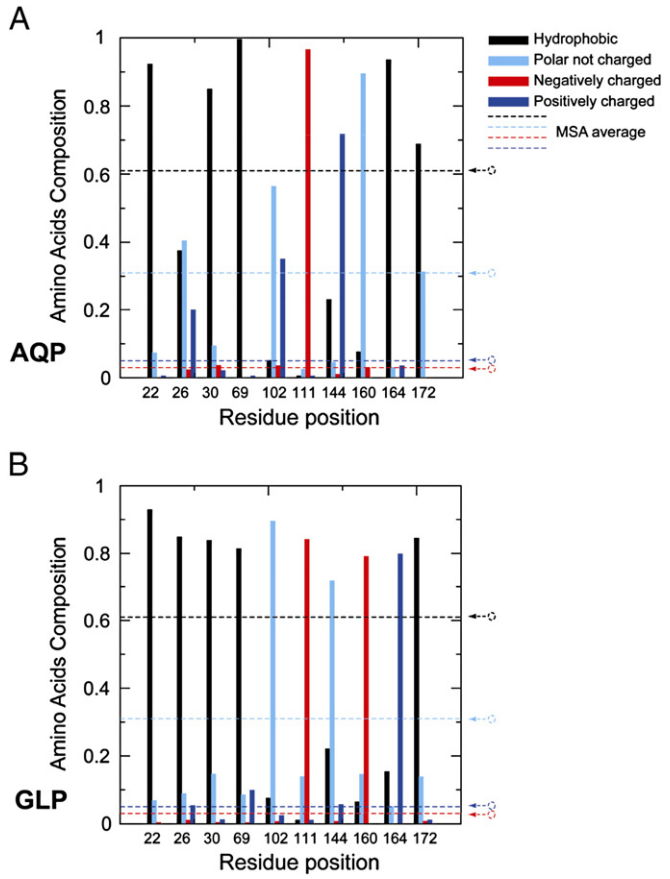

Fig. 4. Distribution of amino acid types. Physico-chemical characteristics (hydrophobic, polar and charged) of the ten key residue positions identified for (A) the water transporter class (AQP) and (B) the glycerol transporters (GLP). The average proportion of these 4 types of amino acids through all the 192 MSA positions is shown as dotted lines. The color codes for each amino acid type is indicated (top right corner).

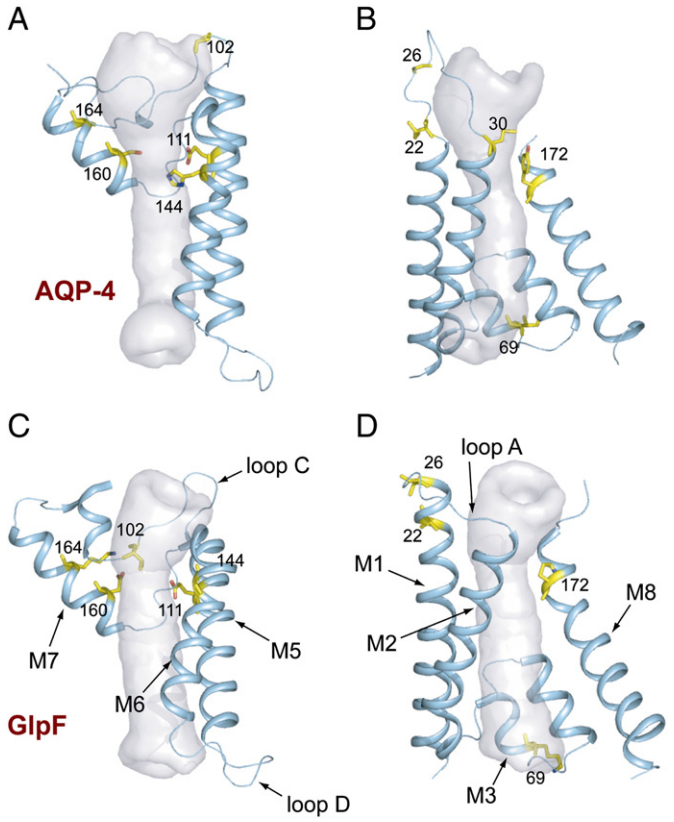

Fig. 5. Map of the main ten positions identified on crystal structures. The ten positions mapped on the crystal structure 3GD8 of a water transporter AQP4 (A–B), and on structure 1LDA of a glycerol facilitator from *E. coli* GlpF (C–D). The positions occupied by mostly polar or charged residues in the MSA are shown on the left (A–C) and the more hydrophobic positions are shown on the right (B–D). The channel opening is shown as a semi-transparent surface. Numbering represents MSA positions. The intracellular side is at the bottom of each panel.

GLP class this position was occupied by the smaller side chain Ala in ~66% of the sequences. Also, at position 172, proline was present in 67% of AQPs, but in GLPs the most common residues were Phe (56%) and Tyr (27%). The importance of these residues may be due to the alteration of the electrostatic potential in the channel, or by the imposition of steric constraints to the passage of molecules. It is therefore reasonable to suspect that these residues may be close to the lining of the pore in the aquaporin monomer. Indeed, this is the case (Fig. 5) for the water transporter human AQP4 (3GD8 [58], panels A–B), or for the *E. coli* glycerol facilitator, GlpF (1LDA [29], panels C–D). All residues corresponding to those MSA positions are located within 9 Å from the lumen of the channel in the two structures (see Table 1), and the percentage of residues within ~5 Å to the lumen is 80% in the AQP4 structure, and 40% in the GlpF structure.

Table 1
The ten key residue positions in the MSA. Equivalence of these positions in GlpF and in AQP4 for reference, and distance to the lining of the channel in both proteins. For structure 1LDA, residue 30 was in a gap in the MSA and therefore it was not included.

Position in MSA	Residue in 1LDA (GlpF)	Residue in 3GD8 (AQP-4)	Distance to channel (Å) 1LDA	Distance to channel (Å) 3GD8
22	30A	57I	5.6	3.7
26	34G	61G	7.8	4.3
30	GAP	70M	–	3.6
69	83K	111I	8.8	1.8
102	137T	152N	2.5	1.7
111	152E	163E	7.8	5.6
144	191G	201H	3.8	1.7
160	207D	217S	5.5	5.1
164	211K	221A	7.2	8.9
172	236P	233Y	7.7	9.6

3.4. Topological location of identified MSA positions

A striking feature of the topological representation of the ten key residues (Fig. 6, dark blue dots) is that only one them (K83 in GlpF) is located in the cytoplasmic half of the protein. The number of key residues obtained depends on the cut-off applied in the analysis of non-redundant classifiers. Therefore, ten more positions were identified with a less stringent cut-off. However, even considering the next ten more relevant MSA positions (37, 43, 44, 50, 100, 136, 151, 154, 156 and 163, equivalent to GlpF residues W48, M54, A55, G61, F135, I183, T198, A201, N203 and P210) most are found in the 'extracellular' half of the protein, in loops A, C and E, C-terminal NPA and ar/R motifs (Fig. 6, light blue dots). Thus, despite the structural arrangement of aquaporins as two hemipores, aquaporins seem to be asymmetrically organized, with an extracellularly oriented face being more involved in substrate selectivity.

3.5. Biological relevance of the ten key residues

The residues in these ten key positions showed couplings to residues in several other positions in the MSA, but for simplicity, only 'internal' couplings, i.e., amongst the 10 key positions are shown in Fig. 7A. Residues located at several of these ten positions in the MSA have already been identified by other authors as crucial for substrate selectivity.

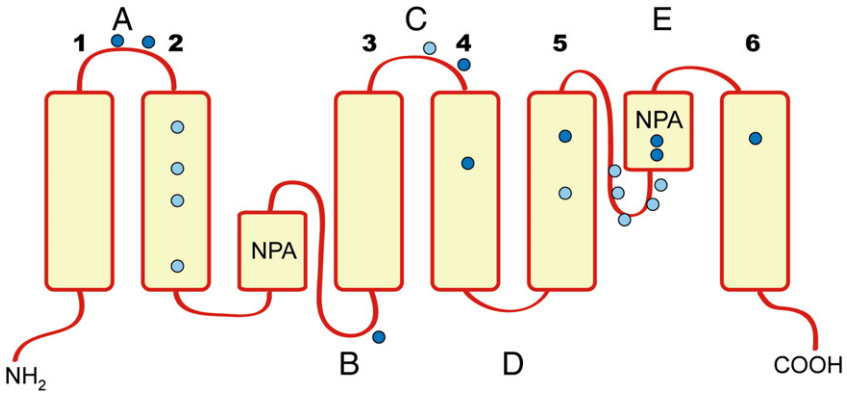

Fig. 6. Topological location of the ten key residues. The ten key positions (Table 1) are indicated as dark blue circles. The next ten most important residues are shown in light blue.

3.5.1. Positions 22, 26 and 30

The functional importance of residues at positions 22, 26 (30A and 34 V in GlpF) and 30 needs to be experimentally confirmed. In GlpF position 30 is not represented, as there is a gap at this position in the alignment (Table 1).

3.5.2. Position 69

Position 69 (83 K in GlpF) is located in loop B and is contiguous to S115 in spinach aquaporin SoPIP2;1. Phosphorylation of S115 increases water transport activity by disrupting interactions between loop D, loop B and the N-terminus [59,60]. Not surprisingly, this position was found to be the most distant to the channel lumen (8.8 Å) in the GlpF crystal structure, 1LDA (see Table 1). Therefore, the reason this residue has been detected in our analysis is unlikely to be due to its contribution to selectivity, via electrostatic or steric effects. Rather, this may obey to the existence of different gating mechanisms present in water channels relative to glycerol transporters. Indeed, phosphorylation, one of the possible mechanisms of aquaporin gating, has been observed in aquaporins 1, 2 4 and 5 (all water transporters [61]), and it takes place at conserved extramembrane serine residues.

Another gating mechanism is through activation by low pH via His protonation (e.g., H193 in loop D in SoPIP2;1 [62]), but we could not identify any position in loop D in our analysis. One possible explanation is that this mechanism of activation is present in members of both types of aquaporins, although not in all of them. The same rationale may also apply to regulation via cation binding, which takes place at cytoplasmic loops, because with the exception of position 69 we could not detect any position in these regions. Indeed, cation regulation has been observed in water channel aquaporins AQP1 and AQP4, but also in the aquaglyceroporin AQP3 [61]. Finally, a side chain conformational change has been proposed as a gating mechanism, (mediated by R189 in *E. coli* Aquaporin Z) [63,64], although the significance of this has not been proved physiologically. This residue (R206 in GlpF) is very well conserved in all aquaporins, and therefore it cannot be detected by our analysis, but it is contiguous to key position 160 (D207 in GlpF, see Table 1). Nevertheless, the fact that it is so well conserved suggests that if it has any regulatory role, it would be common to both types of aquaporins.

3.5.3. Position 102

Position 102 is equivalent to residue T137 in GlpF. This Thr residue is part of the FST triad (FAT in mammals), a highly conserved motif in aquaglyceroporins located close to the pore [36]. A study that compared *E. coli* GlpF (FST motif, higher permeability to glycerol) and an aquaporin from *Plasmodium falciparum* (equal transport of water and glycerol) that contained a WET motif, showed that mutation E125S abolished water permeability [65]. Thus it is likely that both S and T in that motif are important for water/glycerol selectivity, although we could not detect MSA position 101 (corresponding to GlpF S136) in our analysis.

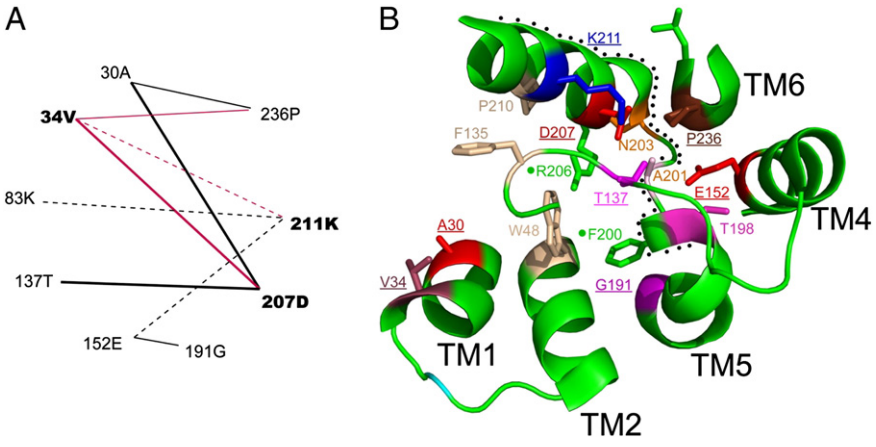

Fig. 7. Couplings and mapping of residues on the structure of GlpF (1LDA). (A) Schematic representation of the couplings observed amongst the ten key positions with reference to GlpF; (B) structure 1LDA of GlpF, showing only the extracellular half. Nine of the ten key residues are underlined. Also shown (not underlined) are six residues from of the next ten, which are also located in that part of the molecule. The important residues F200 and R206, which are part of the ar/R motif, are also shown (green). The TM numbers are indicated, and the row of dots represents the E loop, containing the C-terminal NPA motif.

3.5.4. Positions 111 and 160

Positions 111 and 160 (E152 and D207 in GlpF) correspond to highly conserved pore lining residues belonging to the selectivity filter.

3.5.5. Position 144

Position 144 is equivalent to the lumenally exposed His in water channels human AQP4 and *E. coli* AQP Z (H201 and H174, respectively). In fact, a histidine at that position is a signature of water-selective channels; this residue is directly involved in the selectivity filter, increasing hydrophilicity and also reducing the channel diameter to 1.5 Å in AQP4 [58]. The GLP class contains Gly in ~61% of the sequences at this position. In GlpF (G191), this small residue allows the placement of a phenylalanine (GlpF F200) side chain into the filter that contributes to generate a hydrophobic pocket opposite to the conserved arginine (GlpF R206) [66]. Thus the side chain of F200 substitutes the small threonine and alanine side chains present in Aqp Z and AQP4, respectively.

3.5.6. Positions 160, 164 and 172

In an early sequence analysis [41], five positions were identified corresponding to amino acids conserved in both aquaporins and glycerol channels, which had different physico-chemical properties in the two groups. These positions were named P1–P5 (in human AQP1, T116, S196, A200, F212 and W213, respectively). Three out of these five positions, P2, P3 and P4 are also identified in our current study, and correspond to MSA positions 160, 164 and 172. Position 164 (P3) is equivalent to residue K211 in GlpF, which is in close contact with Ser and Thr sidechains of the FST triad [36]. Mutants of insect aquaporin AQPcic were tested for permeability to water or glycerol, and a mutant in the pair P4, P5 (Y222P/W223L) lost its ability to transport water, acquiring glycerol transport properties [37]. Y222 (P4) in AQPcic corresponds to key residue position 172 in our MSA. Other AQPcic mutants tested at the two residues corresponding to our MSA positions 160 and 164, S205D (P2) and A209K (P3), did not show any increase in glycerol permeability, but a drop in water permeability.

The above experiments showed only partial success in changing substrate specificity and suggest that mutagenesis studies cannot violate existing evolutionary constraints, i.e., any one key position may not be able to modify substrate selectivity if one or more coupled position is not changed simultaneously.

3.6. Prediction of physical contacts in aquaporins

Although it is out of the scope of this manuscript, selecting those coupled pairs *common* to both groups, i.e., water and glycerol transporters, should point to residues important for structural stability, or perhaps common gating mechanisms. In this respect, correlated mutations have been used in the past to predict protein structure, by analyzing patterns of residue replacements in evolutionary trees [67,68]. An accuracy of ~50% was obtained for prediction of physical contacts in water soluble proteins by comparing similarity matrices calculated at each position in an MSA [69]. The latter approach was used to predict physical contacts in 46 diverse aquaporin sequences, by restricting candidate pairs to those expected to be less tan 10 Å apart in the z direction, i.e., normal to the membrane plane [70]. However, many of these pairs did not represent physical contacts. In fact, using seven different algorithms (SCA one of them) to predict contacts in 14 membrane proteins using correlated mutations [71], prediction accuracy was not higher than 50%, and in globular proteins, this value was not higher than 20% [72]. Thus, many co-evolving pairs, whether contributing to functional or structural stabilization, correspond to spatially distant locations. In our study, in contrast, given that all aquaporins are structurally similar, we have attempted to filter out those contributions that are structural, by focusing only on those pairs that are biased towards one or the other aquaporin groups. This results in an enrichment of functionally relevant couplings. We show that most of these residues are close to the functional part of the aquaporin, i.e., the channel lumen, which validates our initial assumption.

4. Conclusion

We have used SCA analysis to find differentially co-evolved pairs in two phylogenetically classified groups of aquaporin sequences to build, iteratively, a number of non-identical binary classifiers. As much as 50% of the elements within these classifiers correspond to only ten key positions in the MSA. The residues at these positions are located near the lumen of the aquaporin pore, and their functional relevance is confirmed in several cases by available functional data. This analysis has also revealed that most of these residues are located in the extracellular half of aquaporin, highlighting the asymmetry of the molecule despite its apparently symmetric organization in two hemipores. Lastly, these results predict that mutagenesis studies to change permeability should be performed in co-evolved pairs, and that unsuccessful attempts using single or not co-evolved residues may be explained by the violation of evolutionary constraints.

Acknowledgments

This material is based on research/work supported in part by the Singapore National Research Foundation under CRP Award No. NRF-CRP4-2008-02. The funding from National Research Foundation through the Environment and Water Industry Programme Office (EWI) on project 0804-IRIS-02 is also acknowledged.

References

[1] A.S. Verkman, Mammalian aquaporins: diverse physiological roles and potential clinical significance, Expert Rev. Mol. Med. 10 (2008) e13.
[2] J.B. Heymann, P. Agre, A. Engel, Progress on the structure and function of aquaporin 1, J. Struct. Biol. 121 (1998) 191–206.
[3] M.B. Gorin, S.B. Yancey, J. Cline, J.P. Revel, J. Horwitz, The major intrinsic protein (MIP) of the bovine lens fiber membrane: characterization and structure based on cDNA cloning, Cell 39 (1984) 49–59.
[4] J.F. Hubert, L. Duchesne, C. Delamarche, A. Vaysse, H. Gueune, C. Raguenes-Nicol, Pore selectivity analysis of an aquaglyceroporin by stopped-flow spectrophotometry on bacterial cell suspensions, Biol. Cell 97 (2005) 675–686.
[5] F. Sidoux-Walter, N. Pettersson, S. Hohmann, The Saccharomyces cerevisiae aquaporin Aqy1 is involved in sporulation, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 17422–17427.
[6] M.J. Borgnia, P. Agre, Reconstitution and functional comparison of purified GlpF and AqpZ, the glycerol and water channels from Escherichia coli, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 2888–2893.
[7] A. Froger, J.P. Rolland, P. Bron, V. Lagree, F. Le Caherec, S. Deschamps, J.F. Hubert, I. Pellerin, D. Thomas, C. Delamarche, Functional characterization of a microbial aquaglyceroporin, Microbiology 147 (2001) 1129–1135.
[8] E.M. Campbell, A. Ball, S. Hoppler, A.S. Bowman, Invertebrate aquaporins: a review, J. Comp. Physiol. 178 (2008) 935–955.
[9] A.K. Mah, K.R. Armstrong, D.S. Chew, J.S. Chu, D.K. Tu, R.C. Johnsen, N. Chen, H.M. Chamberlin, D.L. Baillie, Transcriptional regulation of AQP-8, a Caenorhabditis elegans aquaporin exclusively expressed in the excretory system, by the POU homeobox transcription factor CEH-6, J. Biol. Chem. 282 (2007) 28074–28086.
[10] M. Suzuki, T. Hasegawa, Y. Ogushi, S. Tanaka, Amphibian aquaporins and adaptation to terrestrial environments: a review, Comp. Biochem. Physiol. 148 (2007) 72–81.
[11] K. Ishibashi, S. Hara, S. Kondo, Aquaporin water channels in mammals, Clin. Exp. Nephrol. 13 (2009) 107–117.
[12] A.J. Yool, Functional domains of aquaporin-1: keys to physiology, and targets for drug discovery, Curr. Pharm. Des. 13 (2007) 3212–3221.
[13] M. Yasui, A. Hazama, T.H. Kwon, S. Nielsen, W.B. Guggino, P. Agre, Rapid gating and anion permeability of an intracellular aquaporin, Nature 402 (1999) 184–187.
[14] S.M. Saparov, K. Liu, P. Agre, P. Pohl, Fast and selective ammonia transport by aquaporin-8, J. Biol. Chem. 282 (2007) 5296–5301.
[15] L.M. Holm, T.P. Jahn, A.L. Moller, J.K. Schjoerring, D. Ferri, D.A. Klaerke, T. Zeuthen, NH3 and NH4+ permeability in aquaporin-expressing Xenopus oocytes, Pflugers Arch. 450 (2005) 415–428.
[16] L.S. King, D. Kozono, P. Agre, From structure to disease: the evolving tale of aquaporin biology, Nat. Rev. Mol. Cell Biol. 5 (2004) 687–698.

[17] R. Wysocki, C.C. Chery, D. Wawrzycka, M. Van Hulle, R. Cornelis, J.M. Thevelein, M.J. Tamas, The glycerol channel Fps1p mediates the uptake of arsenite and antimonite in Saccharomyces cerevisiae, Mol. Microbiol. 40 (2001) 1391–1401.
[18] O.I. Sanders, C. Rensing, M. Kuroda, B. Mitra, B.P. Rosen, Antimonite is accumulated by the glycerol facilitator GlpF in Escherichia coli, J. Bacteriol. 179 (1997) 3365–3367.
[19] G.J. Cooper, W.F. Boron, Effect of PCMBS on CO2 permeability of Xenopus oocytes expressing aquaporin 1 or its C189S mutant, Am. J. Physiol. 275 (1998) C1481–C1486.
[20] M. Herrera, N.J. Hong, J.L. Garvin, Aquaporin-1 transports NO across cell membranes, Hypertension 48 (2006) 157–164.
[21] U. Johanson, M. Karlsson, I. Johansson, S. Gustavsson, S. Sjovall, L. Fraysse, A.R. Weig, P. Kjellbom, The complete set of genes encoding major intrinsic proteins in Arabidopsis provides a framework for a new nomenclature for major intrinsic proteins in plants, Plant Physiol. 126 (2001) 1358–1369.
[22] I.I. Ivanov, A.V. Loktyushkin, R.A. Gus'kova, N.S. Vasil'ev, G.E. Fedorov, A.B. Rubin, Oxygen channels of erythrocyte membrane, Doklady 414 (2007) 137–140.
[23] B.L. de Groot, A. Engel, H. Grubmuller, A refined structure of human aquaporin-1, FEBS Lett. 504 (2001) 206–211.
[24] G.M. Preston, T.P. Carroll, W.B. Guggino, P. Agre, Appearance of water channels in Xenopus oocytes expressing red cell CHIP28 protein, Science 256 (1992) 385–387.
[25] T. Gonen, P. Sliz, J. Kistler, Y. Cheng, T. Walz, Aquaporin-0 membrane junctions reveal the structure of a closed water pore, Nature 429 (2004) 193–197.
[26] H. Sui, B.-G. Han, J.K. Lee, P. Walian, B.K. Jap, Structural basis of water-specific transport through the AQP1 water channel, Nature 414 (2001) 872–878.
[27] Y. Hiroaki, K. Tani, A. Kamegawa, N. Gyobu, K. Nishikawa, H. Suzuki, T. Walz, S. Sasaki, K. Mitsuoka, K. Kimura, A. Mizoguchi, Y. Fujiyoshi, Implications of the aquaporin-4 structure on array formation and cell adhesion, J. Mol. Biol. 355 (2006) 628–639.
[28] R. Horsefield, K. Nordén, M. Fellert, A. Backmark, S. Törnroth-Horsefield, A.C. Terwisscha Van Scheltinga, J. Kvassman, P. Kjellbom, U. Johanson, R. Neutze, High-resolution X-ray structure of human aquaporin 5, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 13327–13332.
[29] E. Tajkhorshid, P. Nollert, M.A. Jensen, L.J.W. Miercke, J. O'Connell, R.M. Stroud, K. Schulten, Control of the selectivity of the aquaporin water channel family by global orientational tuning, Science 296 (2002) 525–530.
[30] J.K. Lee, D. Kozono, J. Remis, Y. Kitagawa, P. Agre, R.M. Stroud, Structural basis for conductance by the archaeal aquaporin AqpM at 1.68 Ã…, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 18932–18937.
[31] D.F. Savage, P.F. Egea, Y. Robles-Colmenares, J.D. O'Connell III, R.M. Stroud, Architecture and selectivity in aquaporins: 2.5 a X-ray structure of aquaporin Z, PLoS Biol. 1 (2003) E72.
[32] S. Tornroth-Horsefield, Y. Wang, K. Hedfalk, U. Johanson, M. Karlsson, E. Tajkhorshid, R. Neutze, P. Kjellbom, Structural mechanism of plant aquaporin gating, Nature 439 (2006) 688–694.
[33] R. Zardoya, Phylogeny and evolution of the major intrinsic protein family, Biol. Cell 97 (2005) 397–414.
[34] J.H. Park, M.H. Saier Jr., Phylogenetic characterization of the MIP family of transmembrane channel proteins, J. Membr. Biol. 153 (1996) 171–180.
[35] K. Murata, K. Mitsuoka, T. Hirai, T. Walz, P. Agre, J.B. Heymann, A. Engel, Y. Fujiyoshi, Structural determinants of water permeation through aquaporin-1, Nature 407 (2000) 599–605.
[36] D. Fu, A. Libson, L.J. Miercke, C. Weitzman, P. Nollert, J. Krucinski, R.M. Stroud, Structure of a glycerol-conducting channel and the basis for its selectivity, Science 290 (2000) 481–486.
[37] V. Lagree, A. Froger, S. Deschamps, J.F. Hubert, C. Delamarche, G. Bonnec, D. Thomas, J. Gouranton, I. Pellerin, Switch from an aquaporin to a glycerol channel by two amino acids substitution, J. Biol. Chem. 274 (1999) 6817–6819.
[38] J.S. Hub, B.L. de Groot, Mechanism of selectivity in aquaporins and aquaglyceroporins, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 1198–1203.
[39] I.S. Wallace, D.M. Roberts, Homology modeling of representative subfamilies of Arabidopsis major intrinsic proteins. Classification based on the aromatic/arginine selectivity filter, Plant Physiol. 135 (2004) 1059–1068.
[40] C.M. Krane, D.L. Goldstein, Comparative functional analysis of aquaporins/glyceroporins in mammals and anurans, Mamm. Genome 18 (2007) 452–462.
[41] A. Froger, B. Tallur, D. Thomas, C. Delamarche, Prediction of functional residues in water channels and related proteins, Protein Sci. 7 (1998) 1458–1468.
[42] D.F. Savage, J.D. O'Connell 3rd, L.J. Miercke, J. Finer-Moore, R.M. Stroud, Structural context shapes the aquaporin selectivity filter, Proc Natl Acad Sci U S A 107 (2010) 17164–17169.
[43] S.W. Lockless, R. Ranganathan, Evolutionarily conserved pathways of energetic connectivity in protein families, Science 286 (1999) 295–299.
[44] R.D. Finn, J. Tate, J. Mistry, P.C. Coggill, S.J. Sammut, H.R. Hotz, G. Ceric, K. Forslund, S.R. Eddy, E.L. Sonnhammer, A. Bateman, The Pfam protein families database, Nucleic Acids Res. 36 (2008) D281–D288.
[45] W.P. Russ, D.M. Lowery, P. Mishra, M.B. Yaffe, R. Ranganathan, Natural-like function in artificial WW domains, Nature 437 (2005) 579–583.
[46] M. Socolich, S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner, R. Ranganathan, Evolutionary information for specifying a protein fold, Nature 437 (2005) 512–518.
[47] J.B. Heymann, A. Engel, Aquaporins: phylogeny, structure, and physiology of water channels, News Physiol. Sci. 14 (1999) 187–193.
[48] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[49] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[50] O. Rivoire, A, Ranganathan, A summary of SCA calculations, note 103 (2008).
[51] J.R. Quinlan, C4.5; Programs for Machine Learing, Morgan Kaufmann, San Mateo, CA, 1988.
[52] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed. Wiley, 2002.
[53] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[54] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for the classification and prediction of beta-turn types, J. Pept. Sci. 8 (2002) 297–301.
[55] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect, J. Cell. Biochem. 84 (2002) 343–348.
[56] R. Karchin, K. Karplus, D. Haussler, Classifying G-protein coupled receptors with support vector machines, Bioinformatics 18 (2002) 147–159.
[57] S.L. Lamers, M. Salemi, M.S. McGrath, G.B. Fogel, Prediction of R5, X4, and R5X4 HIV-1 coreceptor usage with evolved neural networks, IEEE/ACM Trans. Comput. Biol. Bioinform. 5 (2008) 291–300.
[58] J.D. Ho, R. Yeh, A. Sandstrom, I. Chorny, W.E. Harries, R.A. Robbins, L.J. Miercke, R.M. Stroud, Crystal structure of human aquaporin 4 at 1.8 A and its mechanism of conductance, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 7437–7442.
[59] M. Nyblom, A. Frick, Y. Wang, M. Ekvall, K. Hallgren, K. Hedfalk, R. Neutze, E. Tajkhorshid, S. Tornroth-Horsefield, Structural and functional analysis of SoPIP2;1 mutants adds insight into plant aquaporin gating, J. Mol. Biol. 387 (2009) 653–668.
[60] I. Johansson, M. Karlsson, V.K. Shukla, M.J. Chrispeels, C. Larsson, P. Kjellbom, Water transport activity of the plasma membrane aquaporin PM28A is regulated by phosphorylation, Plant Cell 10 (1998) 451–459.
[61] K. Hedfalk, S. Tornroth-Horsefield, M. Nyblom, U. Johanson, P. Kjellbom, R. Neutze, Aquaporin gating, Curr. Opin. Struct. Biol. 16 (2006) 447–456.
[62] C. Tournaire-Roux, M. Sutka, H. Javot, E. Gout, P. Gerbeau, D.T. Luu, R. Bligny, C. Maurel, Cytosolic pH regulates root water transport during anoxic stress through gating of aquaporins, Nature 425 (2003) 393–397.
[63] J. Jiang, B.V. Daniels, D. Fu, Crystal structure of AqpZ tetramer reveals two distinct Arg-189 conformations associated with water permeation through the narrowest constriction of the water-conducting channel, J. Biol. Chem. 281 (2006) 454–460.
[64] L. Xin, H. Su, C.H. Nielsen, C. Tang, J. Torres, Y. Mu, Water permeation dynamics of AqpZ: A tale of two states, Biochim Biophys Acta (2011).
[65] E. Beitz, S. Pavlovic-Djuranovic, M. Yasui, P. Agre, J.E. Schultz, Molecular dissection of water and glycerol permeability of the aquaglyceroporin from *Plasmodium falciparum* by mutational analysis, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 1153–1158.
[66] Y. Wang, E. Tajkhorshid, Molecular mechanisms of conduction and selectivity in aquaporin water channels, J. Nutr. 137 (2007) 1509S–1515S.
[67] I.N. Shindyalov, N.A. Kolchanov, C. Sander, Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng. 7 (1994) 349–358.
[68] D. Altschuh, A.M. Lesk, A.C. Bloomer, A. Klug, Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus, J. Mol. Biol. 193 (1987) 693–707.
[69] U. Gobel, C. Sander, R. Schneider, A. Valencia, Correlated mutations and residue contacts in proteins, Proteins 18 (1994) 309–317.
[70] J.B. Heymann, A. Engel, Structural clues in the sequences of the aquaporins, J. Mol. Biol. 295 (2000) 1039–1053.
[71] A. Fuchs, A.J. Martin-Galiano, M. Kalman, S. Fleishman, N. Ben-Tal, D. Frishman, Co-evolving residues in membrane proteins, Bioinformatics 23 (2007) 3312–3319.
[72] A.A. Fodor, R.W. Aldrich, Influence of conservation on calculations of amino acid covariance in multiple sequence alignments, Proteins 56 (2004) 211–221.