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.

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,