8. L. E. Graham, C. F. Delwiche, B. D. Mishler, *Adv. Bryol.* **4**, 213 (1991).
9. B. D. Mishler, S. P. Churchill, *Brittonia* **36**, 406 (1984).
10. L. E. Graham, L. W. Wilcox, *Algae* (Prentice-Hall, Upper Saddle River, NJ, 2000).
11. M. Feist, N. Grambast-Fessard, in *Calcareous Algae and Stromatolites*, R. Riding, Ed. (Springer-Verlag, Berlin, 1991), pp. 189–203.
12. J. Felsenstein, *Syst. Zool.* **27**, 401 (1978).
13. Supplementary material is available on *Science* Online at www.sciencemag.org/cgi/content/full/294/5550/2351/DC1.
14. Polymerase chain reaction (PCR) and sequencing: Total cellular DNA was isolated by the CTAB method [J. J. Doyle, J. L. Doyle, *Phytochem. Bull.* **19**, 11 (1987)], UNSET method (a high-urea, SDS extraction buffer) or using the Nucleon Phytopure Plant DNA extraction kit (Amersham Pharmacia Biotech) following the manufacturer's protocol from fresh thalli growing in uni-algal condition. The genes were amplified by PCR with gene specific primers (*atpB* upstream: 5′-TGTTACTTGTGAAGTTCAACA-3′; *atpB* downstream: 5′-CTAAATAAAATGCTTGTTCAGG-3′; *rbcL* upstream: 5′-ATGTCACCACAAACAGAAACTA-AAGC-3′; *rbcL* downstream: 5′-AATTCAAATTTA-ATTTCTTTCC-3′; *nad5* upstream: 5′-GTAGGT-GATTTTGGATTAGC-3′: *nad5* downstream: 5′-GTACCTAAACCAATCATCATATC-3′; SSU upstream: 5′-GTAGTCATATGCTTGTCTC-3′: SSU downstream: 5′-CTTGTTACGACTTCTCCT-3′) and sequenced using either an ABI-PRISM 377 or 3100 DNA sequencer (PE Applied Biosystems) according to the manufacturer's protocols. The resulting sequence chromatograms were edited and compiled into a single alignment using Sequencher 3.1.1 (Gene Codes Corp.) and exported in NEXUS format for phylogenetic analyses. Many published SSU rRNA gene sequences were difficult to align to published secondary structure models. Small subunit sequences that could not be matched to such structure models were resequenced for this study (*13*). A single intron was found in the *Coleochaete orbicularis nad5* sequence and the distribution of introns in *nad5* was examined in the taxa within our study. No introns were found in any other species of *Coleochaete* or other algal charophyte *nad5* sequence sampled. Introns with the same insertion point as that of *C. orbicularis* were only found in *Sphagnum* (a moss) and *Marchantia* (a liverwort) which share a sequence identity of 69.39%, compared with only 37.82% and 37.81% to *C. orbicularis*, respectively. *Anthoceros* (a hornwort) has an apparently unrelated intron inserted 128 base pairs downstream with 37.35% identity with that of *Sphagnum*, 35.99% identity to *Marchantia*, and 39.46% to *C. orbicularis*. For comparison, pairs of random sequences with similar base composition and length as the natural sequences had an average of 37.78% sequence identity. These data suggest that the *C. orbicularis nad5* intron was acquired independently from that shared by *Sphagnum* and *Marchantia*.
15. J. P. Huelsenbeck, J. P. Bollback, in *Handbook of Statistical Genetics*, M. Bishop, Ed. (Wiley, London, 2001).
16. J. P. Huelsenbeck, F. Ronquist, R. Nielsen, J. P. Bollback, *Science* **294**, 2310 (2001).
17. D. Bhattacharya, K. Weber, S. S. An, W. Berning-Koch, *J. Mol. Evol.* **47**, 544 (1998).
18. H. J. Sluiman, C. Guihal, *J. Phycol.* **35**, 395 (1999).
19. C. Lemieux, C. Otis, M. Turmel, *Nature* **403**, 649 (2000).
20. C. Lemieux, C. Otis, M. Turmel, in press.
21. F. E. Fritsch, *The Structure and the Reproduction of the Algae* (Cambridge Univ. Press, London, 1935), vol. I.
22. J. D. Pickett-Heaps, H. J. Marchant, *Cytobios* **6**, 255 (1972).
23. K. R. Mattox, K. D. Stewart, in *The Systematics of the Green Algae*, D. E. G. Irvine, D. M. John, Eds. (Academic Press, London, 1984), pp. 29–72.
24. B. D. Mishler, S. P. Churchill, *Cladistics* **1**, 305 (1985).
25. Alternative hypotheses that were explored include: *Coleochaete orbicularis* sister to land plants, PP = 0.0, ML = 0.0%; *Coleochaete* sister to land plants, PP = 0.0, ML = 0.0%; Coleochaetales sister to land plants, PP = 0.0, ML = 0.0%; Coleochaetales sister to Charales, PP = 0.0, ML = 0.4%.
26. T. M. Duncan, K. S. Renzaglia, D. J. Garbary, *Pl. Syst. Evol.* **204**, 125 (1997).
27. M. E. Cook, L. E. Graham, C. E. J. Botha, C. A. Lavin, *Am. J. Bot.* **84**, 1169 (1997).
28. M. Feist, R. Feist, *Nature* **385**, 401 (1997).
29. H. Tappan, *The Paleobiology of Plant Protists* (Freeman, New York, 1980).
30. We thank T. Bachvaroff, T. Cooke, G. French, M. Hibbs, J. Lewandowski, T. Marushak, and E. Zimmer for critical comments; C. Drummond, S. Snyder, and A. Zeccardi for technical assistance; J. Bollback and J. Huelsenbeck for important discussions and assistance with Bayesian analyses; M. Casanova, M. Feist, and V. Proctor for material; F. Lang *et al.*, C. Lemieux, C. Otis, and M. Turmel for unpublished sequence data; and S. Fritz, A. Kaspar, R. Sudman, K. Sytsma, and the GPPRGC ("Deep Green"; USDA) for help with development of this project. This work was supported by NSF grant DEB-9978117 and is dedicated to the memory of C. C. Delwiche.

Water Permeation Across Biological Membranes: Mechanism and Dynamics of Aquaporin-1 and GlpF

Bert L. de Groot and Helmut Grubmüller*

"Real time" molecular dynamics simulations of water permeation through human aquaporin-1 (AQP1) and the bacterial glycerol facilitator GlpF are presented. We obtained time-resolved, atomic-resolution models of the permeation mechanism across these highly selective membrane channels. Both proteins act as two-stage filters: Conserved fingerprint [asparagine-proline-alanine (NPA)] motifs form a selectivity-determining region; a second (aromatic/arginine) region is proposed to function as a proton filter. Hydrophobic regions near the NPA motifs are rate-limiting water barriers. In AQP1, a fine-tuned water dipole rotation during passage is essential for water selectivity. In GlpF, a glycerol-mediated "induced fit" gating motion is proposed to generate selectivity for glycerol over water.

Aquaglyceroporins constitute a large family of integral membrane proteins that facilitate highly efficient and specific passive permeation of water and other small uncharged solutes across biological membranes (*1*, *2*). Osmotic water regulation is essential for all life forms, and aquaglyceroporins are found throughout nature, with nearly 300 proteins identified and sequenced so far. In humans, more than 10 different aquaporins with specialized functionality are expressed in tissues as diverse as kidney, red blood cells, and brain. Malfunctions of these proteins cause a wide range of diseases, including nephrogenic diabetes insipidus, congenital cataract, and impaired hearing (*1*, *3*, *4*).

The human water channel aquaporin-1 (AQP1) (Fig. 1) (*5*) permeates water molecules across the membrane at a rate of $3 \times 10^9 \text{ s}^{-1}$ per channel (*6–8*), with an activation energy nearly as low as the one associated with the self-diffusion rate in bulk water (*8*). The homologous bacterial glycerol facilitator GlpF is selective for glycerol and other linear alcohols (*9*, *10*) and shows lower water permeability (*10*, *11*) despite a wider pore. The low activation energies allow one to study entire water-permeation events through both proteins by molecular dynamics (MD) simulations in "real time," without the need to accelerate the process by additional driving forces.

The structural models of human AQP1 (*12–14*) and the atomic structure of GlpF from *Escherichia coli* (*15*) have confirmed and extended the early sequence-based "hourglass" model (*16*): The walls of the pore are formed by six transmembrane helices, 1 through 6, connected by five loops, A through E; the pore center is formed by the two highly conserved fingerprint asparagine-proline-alanine (NPA) motifs contained in the B and E loops, which fold back into the protein. The C-terminal halves of these two loops form two short helices that together form a seventh, kinked transmembrane helix. Despite a wealth of experimental data, major issues need to be resolved at the atomic level: How is this extremely high rate achieved while maintaining strict selectivity? How are ions, and particularly protons, excluded, even though they are known to be conducted well by hydrogen-bonded water chains (*17*, *18*)? What is the exact pathway of water molecules through the channel? How are the known structural differences between AQP1 and GlpF reflected in the permeation mechanism? An especially intriguing question is how GlpF facilitates permeation of (larger) glycerol mole-

Theoretical Molecular Biophysics Group, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany.

*To whom correspondence should be addressed. E-mail: hgrubmu@gwdg.de

cules while hindering passage of (smaller) water molecules.

We have carried out extensive MD simulations (19) of AQP1 and GlpF in their biologically active form (Fig. 1): as tetramers (20, 21), embedded in a fully solvated bilayer membrane. Each of the simulations covered 10 ns. For AQP1, 205 water molecules visited the four pore regions, and 16 full-permeation events were observed, in good agreement with the experimental rate. In both AQP1 and GlpF, no permeation was observed for the central cavity, as has been expected from experiments (16, 22, 23).

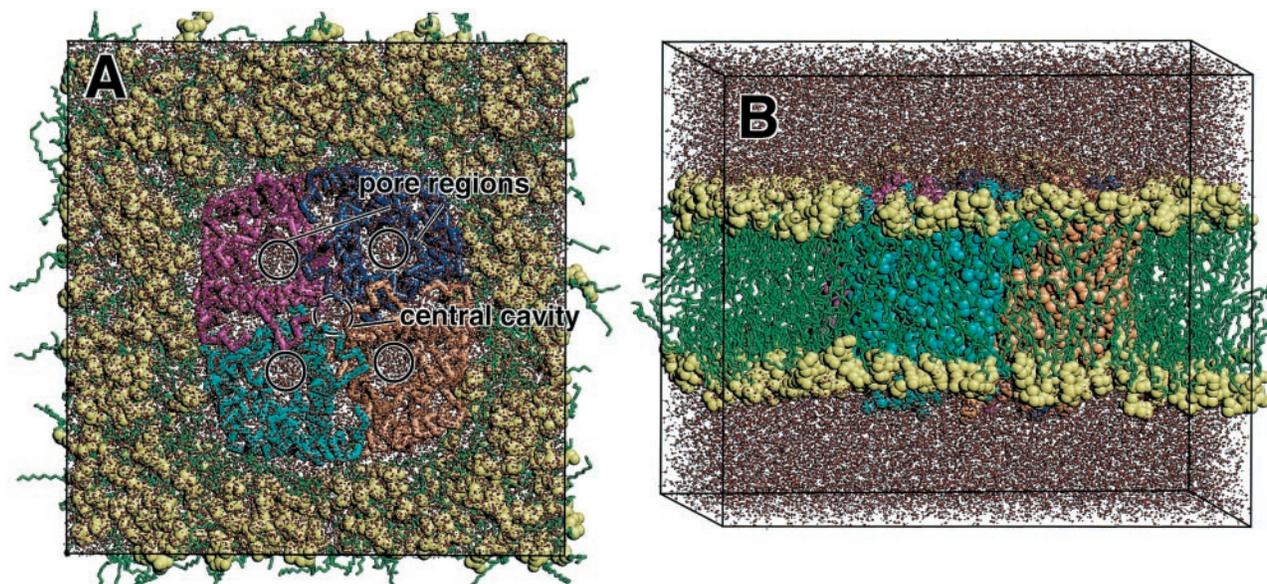Figures 2 and 3 focus on the permeation

Fig. 1. System setup for AQP1 and GlpF simulations. (**A**) Top view and (**B**) side view. All MD simulations were carried out with full electrostatics in a periodic simulation box containing the aquaglyceroporin tetramer (blue, cyan, orange, magenta), embedded within a POPE lipid bilayer (yellow head groups and green tails) surrounded by water (red, white). The total system consists of about 101,000 atoms (19).
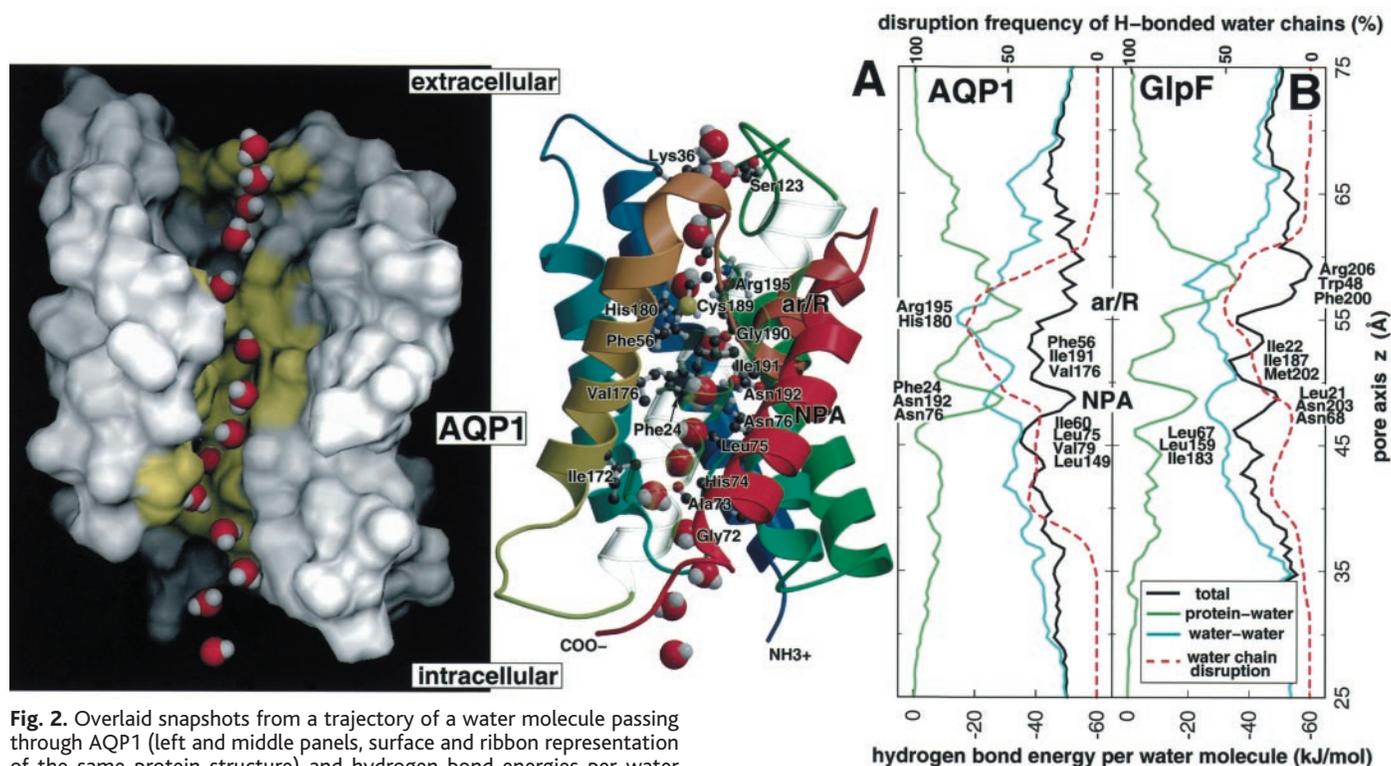

Fig. 2. Overlaid snapshots from a trajectory of a water molecule passing through AQP1 (left and middle panels, surface and ribbon representation of the same protein structure) and hydrogen bond energies per water molecule (right). The permeation event shown on the left lasted 3.3 ns. Those residues that interact with water molecules are colored yellow in the surface representation and are labeled in the ribbon representation, most notably the NPA region (between Asn[76], Asn[192], and Phe[24]) and the ar/R region around Arg[195]. In the surface representation, the front part of the channel has been removed to enable a view at the pore. The right panel shows hydrogen bond energies per water molecule along the pore axis, averaged over all observed (partial) permeation events (205 and 252 for AQP1 and GlpF, respectively). Hydrogen bond energies were calculated from donor-acceptor distances as described (38). The rendered protein model was made with BOBSCRIPT (39) and Raster3D (40).

mechanism. By averaging over the large number of partial permeation events, accurate profiles could be obtained. At the intracellular side (bottom), the pore is relatively wide (see also the red curves in Fig. 3, right), and the protein interacts only weakly with the water molecules (Fig. 2, green curve). The main interaction sites are formed by the backbone carbonyl groups of the residues preceding the first NPA motif, Gly[72] and Ala[73], and the His[74] side chain. Farther up the pathway, the first strong interaction site with water is formed, with both asparagines of the NPA motifs on one side of the pore, and the hydrophobic side chains of Phe[24], Val[176], and Ile[191] on the other. At the extracellular side of the NPA motifs, nearly symmetric to the intracellular side, the carbonyl groups of residues Ile[191], Gly[190], and Cys[189] interact with the water molecules in the pore.

Unexpectedly, 8 Å above the NPA motif, a second main interaction site is found, formed by the aromatic side chains of Phe[56] and His[180], and the positively charged Arg[195] (ar/R). Within this ar/R region, which also forms the narrowest part of the pore (Fig. 3, red curves), the hydrophobic Phe[56] side chain orients the water molecules such as to enforce strong hydrogen bonds to Arg[195] and His[180]. These residues are conserved in the water-selective aquaglyceroporins (24). Farther up the pore, at the extracellular side, the A and C loops interact with water mainly through Lys[36] and Ser[123], respectively.

Contiguous hydrogen-bonded water chains through the channels, a prerequisite for proton conductance, are not observed in any of the MD snapshots for either AQP1 or GlpF. The red curves in Fig. 2 show that for both

proteins, such chains are most frequently interrupted within the ar/R region around the conserved (24) arginine (Arg[195] in AQP1 and Arg[206] in GlpF). This is also the region where water-water hydrogen bonds are weakest (blue curves). Together with the electrostatic repulsion by the positively charged arginine, this characterizes the ar/R region as the main filter for protons and other positive ions, including hydronium ions.

Contrary to what has been expected from more qualitative models (15), our simulations reveal the energetic cost of the disruption of water shells (Fig. 2, blue curves) in the ar/R region to be fully compensated by polar water-protein interactions (green curves). A similar compensation occurs within the NPA regions. Such disruption is thus unlikely to be rate-limiting. Instead, as can also be seen from the free-energy patterns in Fig. 3 (black curves), the simulations identify the hydrophobic zones adjacent to the NPA regions, where such compensation cannot occur, as the rate-limiting barriers of both AQP1 and GlpF.

Focusing on the different specificities of the two channels, the qualitative structure-based model of the fully conserved NPA motifs as important selectivity-determining regions (12) and the relevance of the hydrophobic residue facing the NPA motifs (24–26) are confirmed and explained. The simulations show that in GlpF, the NPA region interacts less strongly with passing water molecules than in AQP1 (Fig. 2, green curves). In AQP1, the hydrophobic Phe[24] side chain intrudes more into the pore and thus forces passing water molecules to interact more strongly with the NPA loops than in GlpF (where this Phe is replaced by Leu[21] and thereby acts as a size-exclusion selectivity filter, hindering the passage of larger molecules such as glycerol). Indeed, preliminary results from a simulation of the Leu[21]→Phe mutant of GlpF confirm this finding, showing a reduced water permeability.

Strong translational and orientational control of the water molecules, involving a fine-tuned dipole inversion, is observed in the pore region, particularly for AQP1 (Fig. 3, left). In both proteins, the water dipoles align with the helix macrodipoles (27) caused by the B and E helices, as speculated for AQP1 (12). The center of the rotation lies near the NPA motifs, where both N_2-termini of the helices meet. As suggested by Berendsen (28), only small molecules with a large dipole moment are likely to be able to follow this inversion fast enough. This effect, therefore, is crucial for selectivity. A similar, although less pronounced, pattern is seen for GlpF. Here, this orientation pattern is broken within the ar/R region, in contrast to AQP1, where the dipoles remain oriented perpendicular to the membrane.

The combined enthalpic (mainly hydrogen bonds, Fig. 2) and entropic (Fig. 3)

Fig. 3. Water dipole orientation (left) and water statistics (right) for AQP1 (top) and GlpF (bottom). Orientation and strength of the water dipole moment were calculated along the pore axis for both proteins, averaged over all (partial) permeation events observed in the simulations. The arrowheads represent the positive ends of the dipoles. The protein structures are colored according to the local electrostatic potential [from negative (red) to positive (blue)], as calculated with DELPHI (41). The arrow sizes and colors (from green to yellow) indicate the size of the average dipole moment per water molecule. The right panel shows relative free energies $G_{PMF}(z)$ (kJ/mol) of water molecules (black curves), the entropic contribution TS to $G(z)$ (green curves) due to restrained rotational freedom of the water molecules (42), pore radii $r(z)$ (Å) (red curves), and the degree $c(z)$ of correlated motion of adjacent water molecules along the pore axes z (blue curves). For GlpF, pore dimensions for the first (solid red curve) and last (dashed curve) nanosecond of the simulation are shown separately. The rendered protein models were made with BOBSCRIPT (39) and Raster3D (40).
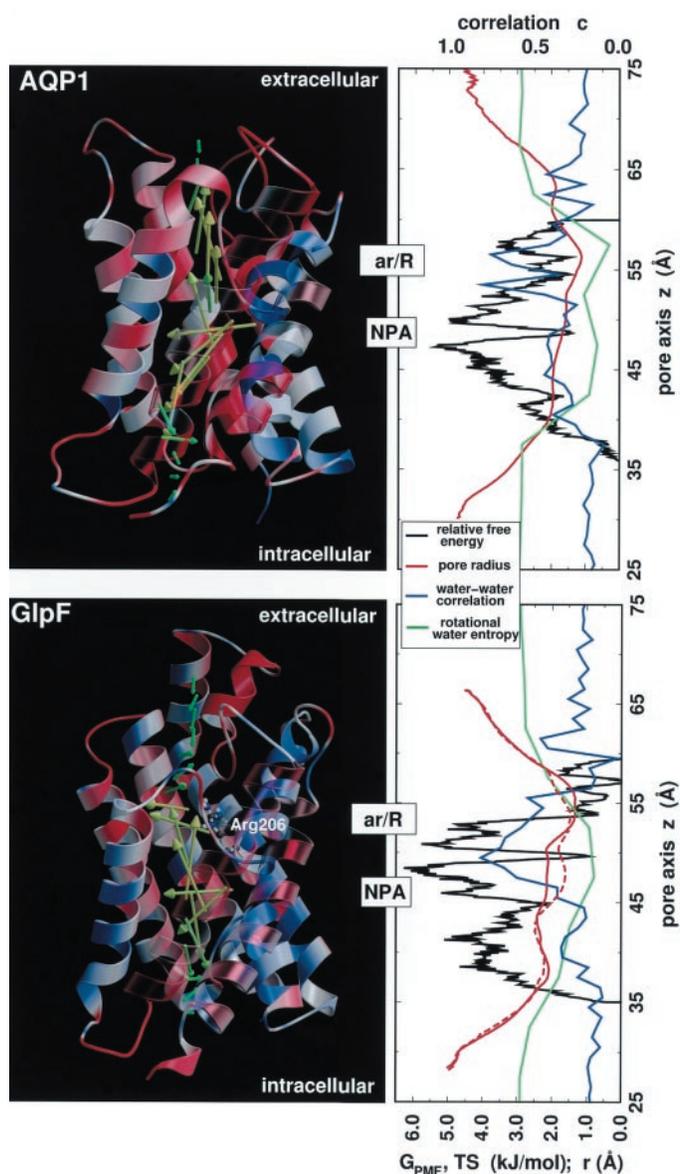
Fig. 4. Schematic summary of the permeation mechanism for AQP1 and GlpF, as extracted from our simulations. In both proteins the ar/R region is the narrowest part of the pore (even more so in AQP1) and forms the upper filter (red). Here, water-water hydrogen bonds are weakened to the largest extent (light blue) as compared with bulk water (darker blue), and therefore, together with the positive charge of the conserved Arg[195] as compared with bulk water (darker blue), and therefore, together with the positive charge of the conserved Arg[195] and Arg[206], renders this region a proton filter. This effect is more pronounced in GlpF than in AQP1. In GlpF, the strongest protein-water hydrogen bonding site (green) is also located in this region, at Arg[206], due to water repulsion by the aromatic Trp[48] and Phe[200] (gray ellipses). Farther down, the second stage of the filter (orange) is located at the conserved NPA motifs and is mainly a size-exclusion selectivity filter. Additionally, the water repulsion due to the hydrophobic Phe[24] group enhances the water-asparagine interactions at the NPA motif and renders this region the strongest protein-water interaction site of AQP1 (green). In both proteins, the dipoles of water molecules passing the pore (red arrows) undergo a rotation during passage. For both AQP1 and GlpF, two rings of exclusively hydrophobic residues, directly adjacent to the NPA regions, form the rate-limiting free-energy barriers (yellow) for water molecules.
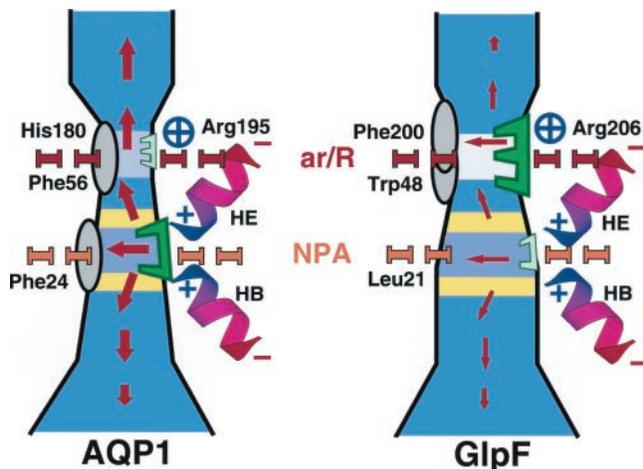
barriers computed for single molecules are much larger than the effective rate-determining free-energy barrier G_{PMF} (Fig. 3, right). This discrepancy suggests that the observed high water-permeation rate is achieved through the highly collective motion of water molecules, which effectively lowers the activation energy. Such strong correlation is actually seen in the simulations (Fig. 3, blue curve). This multiparticle effect renders a straightforward (single-particle) rate estimate from the Arrhenius activation energy, e.g., using Kramers' theory, problematic. Indeed, considerable mismatch is seen between measured rates and corresponding activation energies (*11*).

For GlpF, the water permeability is found to decrease in the course of the simulation from an initial rate (during the time span from 1 to 3 ns) of 2.5×10^9 s^{-1} per monomer to a value of 1.0×10^9 s^{-1} in the last 2 ns. This decrease, which apparently continues at the end of the simulation, is compatible with the experimental rate of 0.5×10^9 s^{-1} (*11*). Also seen is a simultaneous narrowing motion of the pore, localized mainly at the NPA region (Fig. 3, right, red curves) which, however, does not affect the features of the other profiles. We suggest that this motion, which is not seen for AQP1, is a response to the removal of the crystallographic glycerol molecules from the starting structure and is the cause of the observed decrease. This observation challenges the current view (*15*) that the measured low water permeability is primarily caused by disruption of water shells and, instead, suggests "induced fit" gating motions, triggered by glycerol passage, as the main water blockage mechanism of GlpF. Additionally, it confirms the the role of the NPA region as the selectivity filter.

Full functional permeation events have been simulated and analyzed at the atomic level for the prototypic transmembrane water channel AQP1. Biomolecular permeation rates have been computed from first principles, i.e., without relying on the validity of rate theories. The essential features of aquaglyceroporins are multistage filters and task sharing (Fig. 4). For AQP1 and GlpF, the NPA region acts mainly as a size-exclusion selectivity filter, whereas the ar/R region predominantly provides selectivity against ions and protons.

References and Notes

1. P. Agre, M. Bonhivers, M. J. Borgnia, *J. Biol. Chem.* **273**, 14659 (1998).
2. M. Borgnia, S. Nielsen, A. Engel, P. Agre, *Annu. Rev. Biochem.* **68**, 425 (1999).
3. P. M. T. Deen, C. H. van Os, *Curr. Opin. Cell Biol.* **10**, 435 (1998).
4. J. Li, A. S. Verkman, *J. Biol. Chem.* **276**, 31233 (2001).
5. G. M. Preston, T. P. Carroll, W. B. Guggino, P. Agre, *Science* **256**, 385 (1992).
6. M. L. Zeidel, S. V. Ambudkar, B. L. Smith, P. Agre, *Biochemistry* **31**, 7436 (1992).
7. A. N. van Hoek, A. S. Verkman, *J. Biol. Chem.* **267**, 18267 (1992).
8. A. Engel, H. Stahlberg, in *International Review of Cytology*, T. Zeuthen, W. D. Stein, Eds. (Academic Press, New York, in press).
9. K. B. Heller, E. C. Lin, T. H. Wilson, *J. Bacteriol.* **144**, 274 (1980).
10. C. Maurel, J. Reizer, J. I. Schroeder, M. J. Chrispeels, M. H. Saier, *J. Biol. Chem.* **269**, 11869 (1994).
11. M. J. Borgnia, P. Agre, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 2888 (2001).
12. K. Murata *et al.*, *Nature* **407**, 509 (2000).
13. G. Ren, V. S. Reddy, A. Cheng, P. Melnyk, A. K. Mitra, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 1398 (2001).
14. B. L. de Groot, A. Engel, H. Grubmüller, *FEBS Lett.* **504**, 206 (2001).
15. D. Fu *et al.*, *Science* **290**, 481 (2000).
16. J. S. Jung, G. M. Preston, B. L. Smith, W. B. Guggino, P. Agre, *J. Biol. Chem.* **269**, 14648 (1994).
17. R. Pómès, B. Roux, *Biophys. J.* **71**, 19 (1996).
18. P. L. Geissler, C. Dellago, D. Chandler, J. Hutter, M. Parrinello, *Science* **291**, 2121 (2001).
19. As a starting structure for the simulations, the x-ray structure of GlpF and the cryoelectron crystallographic structure of AQP1 were used [Protein Data Bank (PDB) entries 1FX8 and 1H6I, respectively (*14*, *15*)]. The AQP1 structure used (*14*) was significantly improved with respect to the previous structures [PDB entries 1FQY (*12*) and 1IH5 (*13*)], by using the homology to GlpF and extensive crystallographic refinement against cryo-electron microscopic data. Tetramers of both proteins were generated by applying the appropriate symmetry operations. A solvated palmitoyloleoylphosphatidyl-ethanolamine (POPE) bilayer (*29*) containing 340 lipids was equilibrated for 1 ns. Sixty-nine lipids were removed from the bilayer to generate an approximately cubic hole in the membrane. During a short MD simulation, additional forces were applied to expel the remaining carbohydrate tails from the hole. The protein tetramers were then incorporated into the bilayer system, after which the system was resolvated in a bath of SPC water molecules (*30*). The obtained system was allowed to equilibrate for 200 ps with positional restraints on the protein atoms. During this time, anisotropic pressure coupling (*31*) was applied to obtain the correct density in all three spatial dimensions. Chloride ions (four and eight for GlpF and AQP1, respectively) were added to compensate for the net positive charge of the proteins. The total system size comprised 101,449 and 87,644 atoms for GlpF and AQP1, respectively. All MD simulations were carried out with the Gromacs simulation suite (*32*). The Lincs and Settle methods (*33*, *34*) were applied to constrain covalent bond lengths, allowing an integration step of 2 fs. Electrostatic interactions were calculated with the Particle-Mesh Ewald method (*35*). The temperature was kept constant by separately coupling the protein, lipids, and solvent to an external temperature bath (*31*) with a coupling constant of $\tau = 0.1$ ps. The pressure was kept constant by weak ($\tau = 1.0$ ps) coupling in the z direction (normal to the bilayer plane) to a pressure bath (*31*). The gromacs force field was applied. The simulation lengths were 10 ns for each of the two studied proteins. The first nanosecond was considered equilibration, and all analyses were performed over the remaining part. Both proteins were found to be stable in the simulations. The root mean square deviation from the starting (crystal) structure after 2 ns for all backbone atoms calculated over the complete tetramer was as low as 0.151 nm for GlpF and 0.209 nm for AQP1 (*14*). The monomeric, as well as the tetrameric, AQP1 structure, stabilized by monomer-monomer contacts, remained fully intact during the simulations. Moreover, the residues facing the pore, including the NPA regions, were found to be stable throughout the simulations. Restraints as described (*36*) were not necessary.
20. J. M. Verbavatz *et al.*, *J. Cell Biol.* **123**, 605 (1993).
21. T. Braun *et al.*, *EMBO Rep.* **1**, 183 (2000).
22. L. B. Shi, W. R. Skach, A. S. Verkman, *J. Biol. Chem.* **269**, 10417 (1994).
23. Supplementary Web material is available on *Science* Online at www.sciencemag.org/cgi/content/full/294/5550/2353/DC1.
24. J. B. Heymann, A. Engel, *J. Mol. Biol.* **295**, 1039 (2000).
25. A. Engel, Y. Fujiyoshi, P. Agre, *EMBO J.* **19**, 800 (2000).
26. H. Stahlberg *et al.*, *J. Struct. Biol.* **132**, 133 (2000).
27. W. G. J. Hol, P. T. van Duijnen, H. J. C. Berendsen, *Nature* **273**, 443 (1978).
28. H. J. C. Berendsen, personal communication.
29. D. P. Tieleman, H. J. C. Berendsen, *Biophys. J.* **74**, 2786 (1998).
30. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, J. Hermans, in *Intermolecular Forces*, B. Pullman, Ed. (Reidel, Dordrecht, Netherlands, 1981), pp. 331–342.
31. H. J. C. Berendsen, J. P. M. Postma, A. DiNola, J. R. Haak, *J. Chem. Phys.* **81**, 3684 (1984).
32. H. J. C. Berendsen, D. van der Spoel, R. van Drunen, *Comp. Phys. Commun.* **91**, 43 (1995).
33. B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, *J. Comp. Chem.* **18**, 1463 (1997).

34. S. Miyamoto, P. A. Kollman, *J. Comp. Chem.* **13**, 952 (1992).
35. T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **98**, 10089 (1993).
36. F. Zhu, E. Tajkhorshid, K. Schulten, *FEBS Lett.* **504**, 212 (2001).
37. O. S. Smart, J. G. Neduvelil, X. Wang, B. A. Wallace, M. S. P. Sansom, *J. Mol. Graphics* **14**, 354 (1996).
38. E. Espinosa, E. Molins, C. Lecomte, *Chem. Phys. Lett.* **285**, 170 (1998).
39. R. M. Esnouf, *J. Mol. Graphics Model.* **15**, 132 (1997).
40. E. A. Merritt, D. J. Bacon, *Methods Enzymol.* **277**, 505 (1997).
41. I. Klapper, R. Hagstrom, R. Fine, K. Sharp, B. Honig, *Proteins Struct. Funct. Genet.* **1**, 47 (1986).
42. Free-energy profiles $G_{PMF}(z) = -k_B T \ln n(z)$ (PMF,

Genome-Wide Distribution of ORC and MCM Proteins in *S. cerevisiae*: High-Resolution Mapping of Replication Origins

John J. Wyrick,[1,2]* **Jennifer G. Aparicio,**[3]* **Ting Chen,**[3]
John D. Barnett,[2] **Ezra G. Jennings,**[1,2] **Richard A. Young,**[1,2]
Stephen P. Bell,[1,4] **Oscar M. Aparicio**[3]†

DNA replication origins are fundamental to chromosome organization and duplication, but understanding of these elements is limited because only a small fraction of these sites have been identified in eukaryotic genomes. Origin Recognition Complex (ORC) and minichromosome maintenance (MCM) proteins form prereplicative complexes at origins of replication. Using these proteins as molecular landmarks for origins, we identified ORC- and MCM-bound sites throughout the yeast genome. Four hundred twenty-nine sites in the yeast genome were predicted to contain replication origins, and ~80% of the loci identified on chromosome X demonstrated origin function. A substantial fraction of the predicted origins are associated with repetitive DNA sequences, including subtelomeric elements (X and Y') and transposable element–associated sequences (long terminal repeats). These findings identify the global set of yeast replication origins and open avenues of investigation into the role(s) ORC and MCM proteins play in chromosomal architecture and dynamics.

In eukaryotic cells, chromosome replication initiates from DNA loci called origins of replication that are distributed along chromosomes. In *Saccharomyces cerevisiae*, a DNA sequence that functions as an origin is termed an autonomously replicating sequence (ARS) (*1–3*). An 11–base pair (bp) ARS consensus sequence (ACS) is essential for replication initiation and is recognized by the eukaryotic replication initiator, the Origin Recognition Complex (ORC)

(*4–7*). Additional sequences, including an A/T-rich region, are required for origin function (*7, 8*). Although the presence of a match to the ACS is required for ARS activity, it is not sufficient, and the majority of matches to the ACS in the genome do not have ARS activity (*9*). Furthermore, ARS activity varies depending on chromosomal position, suggesting that local chromatin structure influences ARS function (*1*). As a result of these properties, the chromosome-wide identification of origins of replication has been a labor-intensive task (*10*). To date, the ARSs on only 2 of the 16 chromosomes (III and VI), representing about 5% of the genome, have been mapped at high resolution and their chromosomal activities characterized (*11–15*).

Initiation of DNA replication is regulated through the ORC-dependent recruitment of minichromosome maintenance (MCM) proteins to the ORC-origin complexes during G_1 phase of the cell cycle (*16*). As the binding of ORC

and MCM proteins occurs at or very near the origin, we determined the genome-wide locations of ORC- and MCM-binding sites to identify the positions of potential DNA replication origins across the *S. cerevisiae* genome. Chromatin immunoprecipitation (ChIP) was used to enrich for protein-bound DNA that was labeled and hybridized to DNA microarrays in triplicate (*17*). In this study, we used DNA microarrays that included probes to most yeast intergenic sequences and open reading frames (ORFs) for a total of 12,158 loci (*18*). Analysis of the hybridization data generated an average binding ratio (fluorescent intensity of enriched versus unenriched DNA) along with a confidence interval for binding (*P* value) of the protein of interest with each DNA sequence present on the arrays (*17*). We reasoned that loci exhibiting association with both ORC and MCM proteins would represent sites of bona fide origins of replication. We performed five independent experiments to identify the binding sites of Orc1p, Mcm3p, Mcm4p, and Mcm7p and the entire ORC complex (using a polyclonal antibody recognizing all six ORC subunits, α-Orc1-6p) (*18*).

We identified 707 binding sites for Mcm3p, 719 for Mcm4p, 671 for Mcm7p, 568 for Orc1p, and 531 for Orc1-6p (each at $P \leq 0.025$) (*18*). Extensive overlap occurred among these binding sites, particularly among those associated with the MCM proteins. For example, 443 sites bound all three of the MCM proteins, and 206 additional sites bound two of the three MCM proteins (these 649 MCM-binding sites represent 477 nonadjacent sites). Only a few binding sites for each MCM protein (42, 73, and 46 for Mcm3p, Mcm4p, and Mcm7p, respectively) resided at isolated loci (defined as loci where neither the identified locus nor immediately adjacent loci exhibited binding to any of the other proteins tested). Seventy-five sites that bound all three MCM proteins ($P < 0.025$) failed to show ORC binding in either experiment ($P < 0.10$), and 12 sites that bound ORC in both experiments ($P < 0.025$) failed to bind any MCM proteins ($P < 0.10$). The majority of genomic sites tested (90%) did not show ORC or MCM association in any of the experiments.

To determine whether the identified loci

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA. [3]Program in Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089–1340, USA. [4]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: oaparici@usc.edu