D284–D288 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh016

TransportDB: a relational database of cellular membrane transport systems

Qinghu Ren, Katherine H. Kang and Ian T. Paulsen*

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received August 5, 2003; Revised and Accepted September 3, 2003

ABSTRACT

TransportDB (http://www.membranetransport.org) is a relational database designed for describing the predicted cellular membrane transport proteins in organisms whose complete genome sequences are available. For each organism, the complete set of membrane transport systems was identified and classified into different types and families according to putative membrane topology, protein family, bioenergetics and substrate specificities. Web pages were created to provide user-friendly interfaces to easily access, query and download the data. Additional features, such as a BLAST search tool against known transporter protein sequences, comparison of transport systems from different organisms and phylogenetic trees of individual transporter families are also provided. TransportDB will be regularly updated with data obtained from newly sequenced genomes.

INTRODUCTION

Transport systems, which function in the translocation of solutes, play essential roles in cellular metabolism and activities. They mediate the entry of nutrients into cytoplasm and the extrusion of metabolite wastes, maintain a stable internal environment inside the cell by regulating the uptake and efflux of ions, protect cells from environmental insults, and enhance communications between cells through the secretion of proteins, carbohydrates and lipids. Different transport systems differ in their putative membrane topology, energy coupling mechanism and substrate specificities (1). The most common energy coupling mechanisms are the utilization of adenosine triphosphate (ATP), phosphoenolpyruvate (PEP), or chemiosmotic energy in the form of sodium ion or proton electrochemical gradients. The Transporter Classification (TC) system (http://www-biology.ucsd.edu/~msaier/transport/) represents a systematic approach to classify transport systems according to the mode of transport, energy coupling mechanism, molecular phylogeny and substrate specificity (2–4). The transport mode and the energy coupling mechanism serve as the primary base for the classification due to their relatively stable characteristics. There are four characterized classes of solute transporters in the TC system: channels, secondary transporters, primary active transporters and group translocators. Transporters of unknown mechanism or function are included as a distinct class. Channels are energy-independent transporters that exhibit higher rates of transport and lower stereospecificity compared with other transporter classes. Primary active transporters couple the transport process to a primary source of energy, such as a chemical reaction (e.g. ATP hydrolysis). Secondary transporters utilize an ion or solute electrochemical gradient, e.g. proton/sodium motive force, to drive the transport process. Group translocators modify their substrates during the transport process. For example, the bacterial phsphotransferase system (PTS) phosphorylates its sugar substrates using PEP as the phosphoryl donor and energy source and releases them into cytoplasm as sugar–phosphates. Each transporter class is further classified into individual families or superfamilies according to their function, phylogeny and/or substrate specificity (1).

Since the advent of genomic sequencing technologies, such as whole-genome shotgun sequencing, the complete sequences of 135 prokaryotic and eukaryotic genomes have been published to date, with more than 500 additional genome sequencing projects currently underway around the world (Gold Genomes Online Database, http://ergo.integratedgenomics.com/GOLD/). Convenient and effective methods have to be developed to handle and analyze the immense amount of data generated by whole-genome sequencing projects. It has been found that 5–12% of the complete bacterial genome is often dedicated to transport proteins and associated factors (4,5). An in-depth look at transport proteins is vital to the understanding of the metabolic capability of organisms. However, due to the occurrence of large complex transporter gene families, such as the ATP-binding cassette (ABC) and major facilitator superfamily (MFS), and the presence of multiple transporter gene paralogs in many organisms, it is often problematic to annotate these transport proteins by current primary annotation methods. We have been working on a systematic genome-wide analysis of cellular transport systems. Previously, we reported a comprehensive analysis of the transport systems in 18 prokaryotic organisms (4,5) and in yeast (6) based on the TC system. Here we have expanded our analyses to 121 prokaryotic and eukaryotic systems. TransportDB (http://www.membranetransport.org/), a web-integrated database, was built up to store the results of our analyses and to provide user-friendly interfaces to access the data. A semi-automated pipeline was also set up to facilitate the efficient analyses of transport systems.

*To whom correspondence should be addressed. Tel: +1 301 838 3531; Fax: +1 301 838 0200; Email: ipaulsen@tigr.org

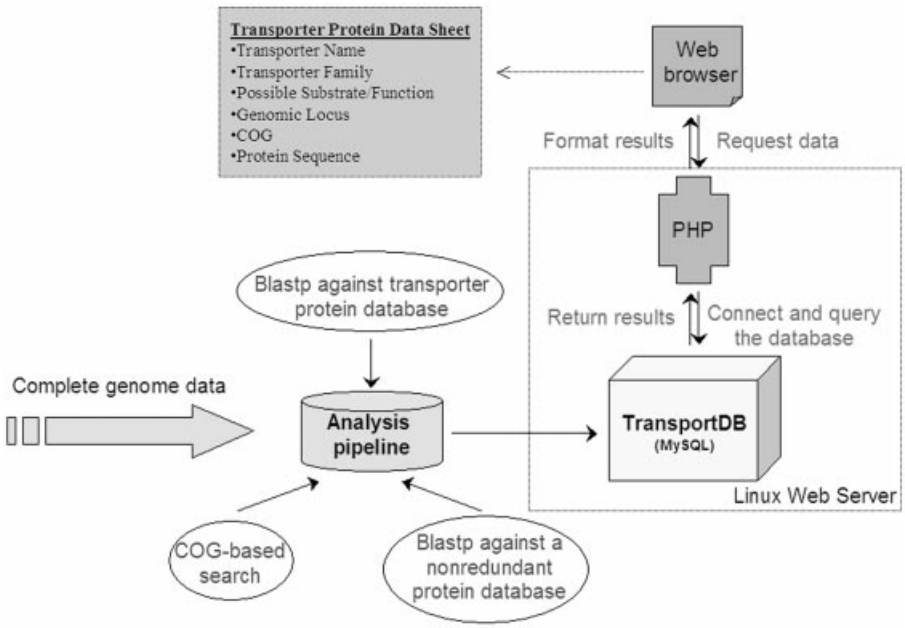

Figure 1. Overview of TransportDB and transport system analysis pipeline. Transporter protein data sets are stored in a MySQL relational database. Users can search the database through a web interface. All web pages are generated dynamically using PHP, which connects, queries the database and formats the results to generate a final data sheet. A pipeline was set up to use the complete genomic protein sequence as input, retrieve transporter proteins and assign them to specific transporter families and substrate/function information. The output can be loaded directly into the database and visualized on the web pages.

DATABASE CONTENT AND STRUCTURE

TransportDB is a MySQL database (http://www.mysql.com/) that is queried using PHP (http://www.php.net/) (see Fig. 1). PHP, a server-side scripting language, mediates the interaction with the user, the database and the computational tools. The database and PHP pages are stored on a Linux web server. TransportDB contains the complete predicted transport profile for each organism, including information on transporter family, TC classification, transporter name, possible substrate/function, genomic locus, COG classification and protein sequence. Where appropriate, links are provided to other databases, such as Entrez (7), COG (8), PubMed (9), TCDB (1) and individual organism genome sequence databases.

Currently, TransportDB contains data from 121 organisms, including 97 bacteria, 16 archaea and eight eukaryota. This collection of organisms represents a broad phylogenetic diversity. A total of 36 137 transporter proteins was assigned to 136 families. Some of these families are very large superfamilies with over a thousand members, such as the ABC superfamily (17 209 total) (10,11), the MFS superfamily (3635 total) (12,13) and the bacterial sugar-specific PTS superfamily (1341 total) (14,15). The transporter profiles from other organisms whose genomic sequencing are underway will be added to the database once their genome sequences are published.

DATABASE ACCESS

TransportDB is available on the web at http://www.membranetransport.org/ (Fig. 2).

The database can be browsed by organism name using the drop-down boxes on the left of the web page. For each organism, its complete membrane transport complement was classified into different families according to the TC classification system. These families were grouped into five distinct types based on mode of transport and energy-coupling mechanisms: ion channels, secondary transporters, ATP-dependent (primary active) transporters, PTSs and unclassified transporters, which have unknown mechanisms of action. Individual transporter types can be accessed by clicking the tabs at the top or the links on the summary page (Fig. 2). For each transporter family, a detailed list of transporters with their predicted substrates is shown with links to the individual protein page which contains genomic locus, COG, protein sequence and annotation information. A summary page is also available for each organism, summarizing the whole transporter system, including transporter types and individual transporter families, and their statistics. TransportDB is searchable by transporter type, transporter family, transporter protein name or substrate. The results are grouped by transporter family and organism, with links to individual family and protein pages.

Comparisons of the transporter contents of different organisms can provide insight into their physiology and life-style. To view the transport profiles across species, we created a 'Compare Organisms' section. Users can choose any two or more organisms to compare their overall numbers of recognized transporters, numbers of transporters relative to genome size and the constituents of each transporter type and family. Previous studies have shown that transporters of similar function characteristically cluster together in

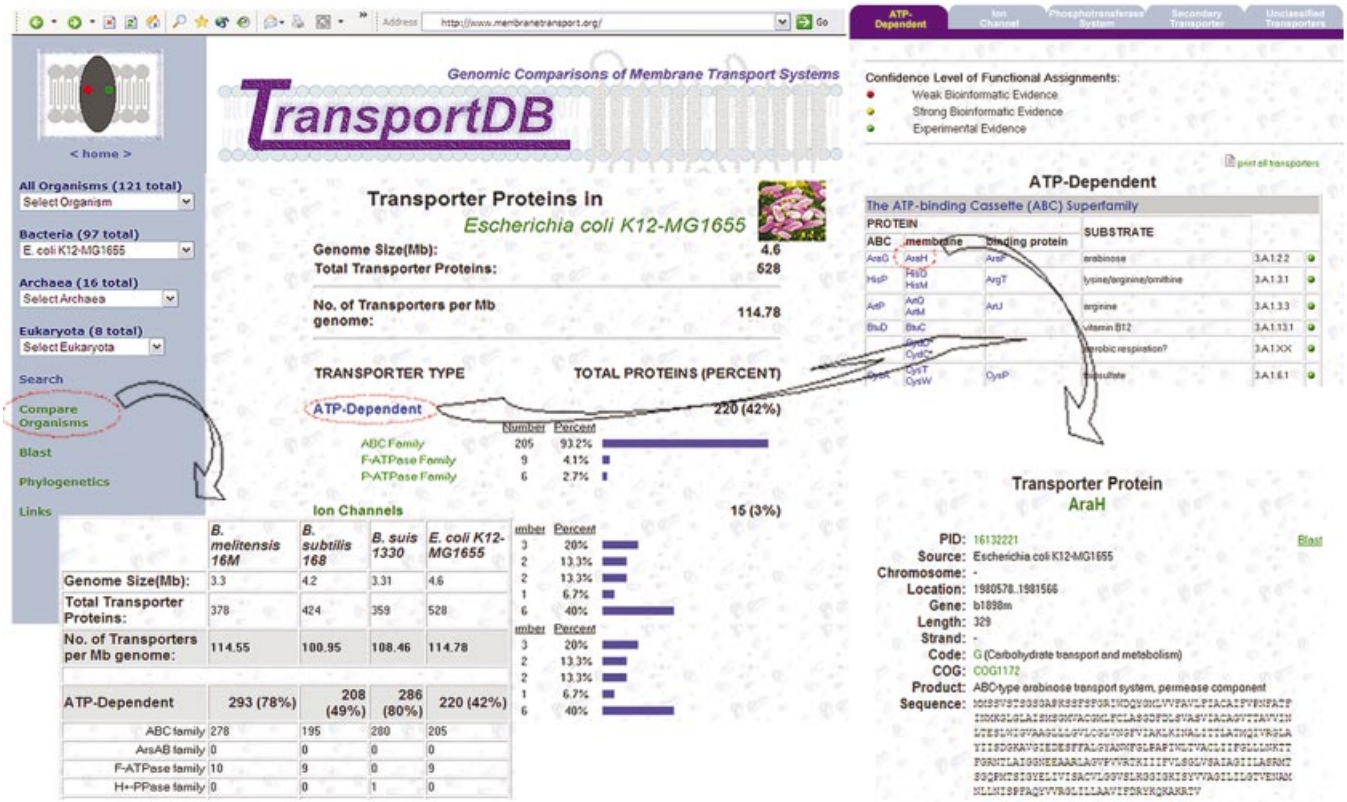

Figure 2. Graphic illustration of the structure of TransportDB. Transport proteins are grouped by organism, transporter type and transporter family. Users can choose an organism from the drop-down boxes at the left. Each organism has a summary page to overview the whole transport system. Individual transporter type or family can be viewed by clicking the links or tabs at the top. Each transporter protein also has an individual page to show the genomic locus, COG information and protein sequence. Links to GenBank, Entrez and COG are also provided. The whole transport profile can also be compared with any other organisms stored in the database.

phylogenetic analyses, hence substrate specificity appears to be a conserved evolutionary trait in transporters (4,5,12,16). In the 'Phylogenetics' section of the TransportDB website, pre-computed neighbor-joining trees for each of the transporter families are available to view. All members of each family in the current database are also available to view or download in FASTA or multiple sequence alignment formats. In addition, the whole transporter database is also available for BLAST search. Users can submit the unknown protein sequence in the 'Blast' section. The output of the BLAST search includes transporter family information in addition to the standard features (17).

It should be noted that TransportDB focuses on solute and ion transport across the cytoplasmic membrane, and hence does not include some types of transporters that are shown in TC classification: outer membrane transporter proteins (18); proteins in the *Escherichia coli* TonB/ExbB/ExbD complex that transduce energy to drive outer membrane transport processes (19); proteins involved in the protein secretory pathways (20); proteins involved in proton and sodium ion-translocating electron transfer processes (21); sodium ion-transporting carboxylic acid decarboxylases (22); flagellar motor proteins (23); proteins involved in DNA uptake (24). Auxiliary transport proteins (such as the MFP family) (25) or membrane–periplasmic auxiliary proteins of the MPA1 and MPA2 families (26) were treated as components of the transporters with which they function, rather than separate transport systems.

ANALYSIS PIPELINE

With the rapid increase in the number of published genomes, efficient and effective approaches are required to speed up transport system analysis processes. Previous methods used by us for transporter analysis (4,5) required intensive personal involvement and manual curation. Recently we have developed a new semi-automated pipeline to analyze a genome-wide transport system, input the data into TransportDB and visualize it on the web page (Fig. 1).

The methodology we have developed is as follows: the complete protein sequences from specific organisms were first searched against the curated set of proteins with family assignment in our transporter protein database for similarity to known or putative transporter proteins using BLAST (27,28). All the proteins with an e-value of <0.001 were collected and searched against a non-redundant general protein database. A web-based interface was created to incorporate the output of two BLAST searches (Fig. 3) and to help a human annotator make a decision and assign possible substrates or functions. The useful information includes: number of hits to the

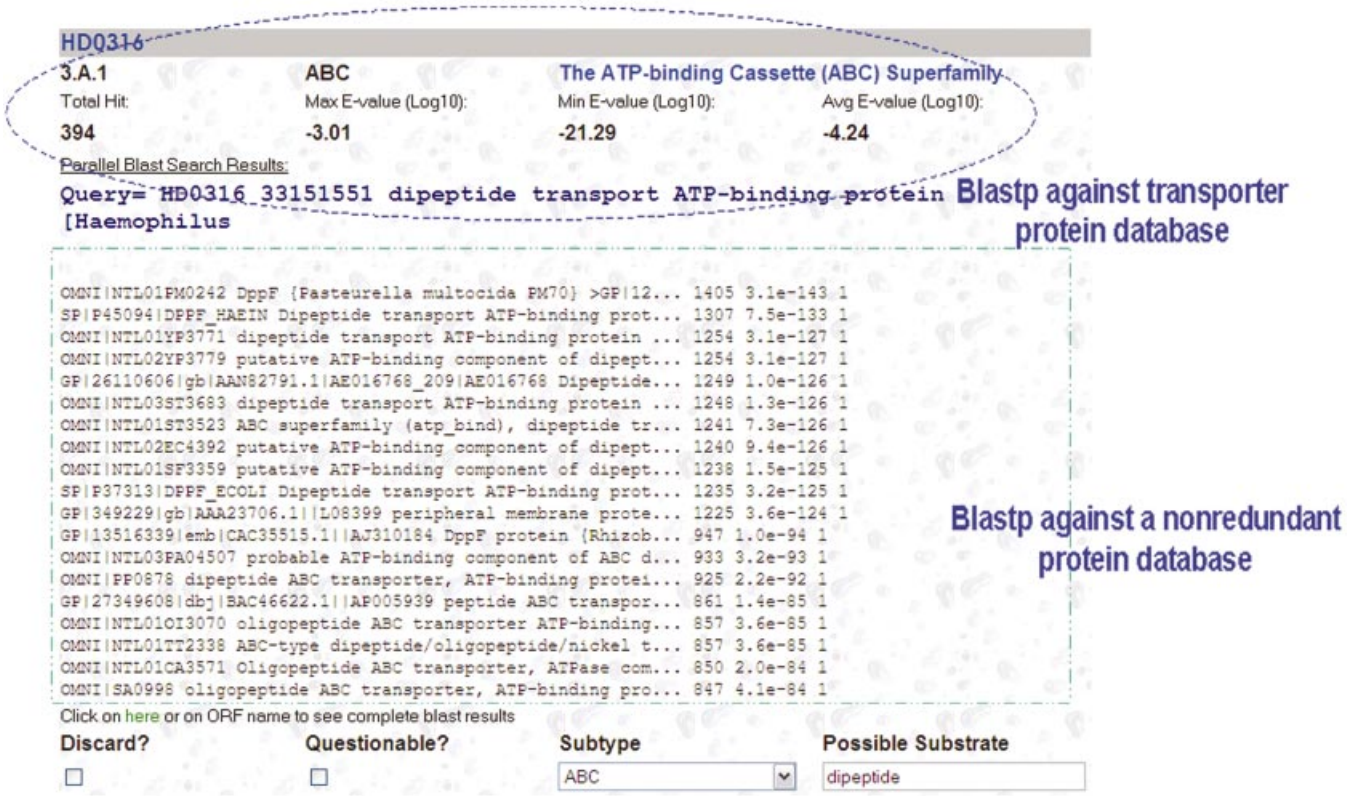

Figure 3. Transporter annotation page. The complete genomic protein sequences were searched against our transporter database and a non-redundant general protein database using BLAST. The results were incorporated into a web-based interface to help the annotator to make a decision on family and substrate properties. Useful information includes: number of hits to the transporter database; maximum, minimum and average e-values; and the description of top hits to the general protein database. Links to TCDB, Entrez and COG are also provided.

transporter database; maximum, minimum and average e-values; and the description of top hits to the general protein database. We also set up direct links between TC family and the COG classification (8) so that COG-based search can also help the annotation processes. The output of the analysis process is in a tab-delimited format, which can be loaded directly into TransportDB and shown on the web pages.

To test the new analysis pipeline, we compared the analysis process on several test genomes by the new pipeline to that by the approaches we had used earlier (4,5). The new pipeline greatly reduced the time annotators spent on the analysis process. In addition, the new pipeline has shown improved sensitivity and selectivity over other approaches. This pipeline has been used in the analysis of over 40 prokaryotic and eukaryotic genomes since its inception.

FUTURE PERSPECTIVES

In summary, we developed a relational database and an analysis pipeline for the comprehensive representation of cellular membrane transport systems in various prokaryotic and eukaryotic organisms. User-friendly web interfaces were designed to easily query the database and access the various features. To our knowledge, this is the only database devoted to the identification and classification of transporter homologs in complete genomes, as well as providing comparative and phylogenetic tools for analyzing the data.

We are continuing to expand the TransportDB database to incorporate data from newly published genomes. TransportDB will be routinely updated at least once per month to ensure the timely report of data. Future planned improvements will include the prediction of transmembrane segments (TMSs) in each transporter protein, prediction of orthologs from different organisms, automated pictorial representation of transport system, links to Swiss-Prot (29) and other online resources, and web-based data submission for TransportDB users.

ACKNOWLEDGEMENTS

We would like to thank Dr Jean-Francois Tomb for comments and suggestions to our web site and Dr Martin Wu for help with the construction of phylogenetic trees. The Linux web server and technical support were kindly provided by Dr Jon Oliver.

REFERENCES

1. Saier,M.H.,Jr (2000) A functional–phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.*, **64**, 354–411.
2. Saier,M.H.,Jr (1999) Genome archeology leading to the characterization and classification of transport proteins. *Curr. Opin. Microbiol.*, **2**, 555–561.
3. Saier,M.H.,Jr (1999) Classification of transmembrane transport systems in living organisms. In VanWinkle,L. (ed.), *Biomembrane Transport*. Academic Press, San Diego, CA, pp. 265–276.

4. Paulsen,I.T., Sliwinski,M.K. and Saier,M.H.,Jr (1998) Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.*, **277**, 573–592.
5. Paulsen,I.T., Nguyen,L., Sliwinski,M.K., Rabus,R. and Saier,M.H.,Jr (2000) Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.*, **301**, 75–100.
6. Paulsen,I.T., Sliwinski,M.K., Nelissen,B., Goffeau,A. and Saier,M.H.,Jr (1998) Unified inventory of established and putative transporters encoded within the complete genome of *Saccharomyces cerevisiae*. *FEBS Lett.*, **430**, 116–125.
7. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
8. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
9. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
10. Tomii,K. and Kanehisa,M. (1998) A comparative analysis of ABC transporters in complete microbial genomes. *Genome Res.*, **8**, 1048–1059.
11. Saurin,W., Hofnung,M. and Dassa,E. (1999) Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. *J. Mol. Evol.*, **48**, 22–41.
12. Pao,S.S., Paulsen,I.T. and Saier,M.H.,Jr (1998) Major facilitator superfamily. *Microbiol. Mol. Biol. Rev.*, **62**, 1–34.
13. Saier,M.H.,Jr, Beatty,J.T., Goffeau,A., Harley,K.T., Heijne,W.H., Huang,S.C., Jack,D.L., Jahn,P.S., Lew,K., Liu,J. *et al*. (1999) The major facilitator superfamily. *J. Mol. Microbiol. Biotechnol.*, **1**, 257–279.
14. Hu,K.Y. and Saier,M.H.,Jr (2002) Phylogeny of phosphoryl transfer proteins of the phosphoenolpyruvate-dependent sugar-transporting phosphotransferase system. *Res. Microbiol.*, **153**, 405–415.
15. Reizer,J., Bachem,S., Reizer,A., Arnaud,M., Saier,M.H.,Jr and Stulke,J. (1999) Novel phosphotransferase system genes revealed by genome analysis—the complete complement of PTS proteins encoded within the genome of *Bacillus subtilis*. *Microbiology*, **145**, 3419–3429.
16. Jack,D.L., Paulsen,I.T. and Saier,M.H. (2000) The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. *Microbiology*, **146**, 1797–1814.
17. Yuan,Y.P., Eulenstein,O., Vingron,M. and Bork,P. (1998) Towards detection of orthologues in sequence databases. *Bioinformatics*, **14**, 285–289.
18. Jeanteur,D., Lakey,J.H. and Pattus,F. (1991) The bacterial porin superfamily: sequence alignment and structure prediction. *Mol. Microbiol.*, **5**, 2153–2164.
19. Braun,V., Pilsl,H. and Gross,P. (1994) Colicins: structures, modes of action, transfer through membranes and evolution. *Arch. Microbiol.*, **161**, 199–206.
20. Saier,M.H.,Jr, Werner,P.K. and Muller,M. (1989) Insertion of proteins into bacterial membranes: mechanism, characteristics and comparisons with the eucaryotic process. *Microbiol. Rev.*, **53**, 333–366.
21. Dimroth,P. (1997) Primary sodium ion translocating enzymes. *Biochim. Biophys. Acta*, **1318**, 11–51.
22. Buckel,W. (2001) Sodium ion-translocating decarboxylases. *Biochim. Biophys. Acta*, **1505**, 15–27.
23. Nguyen,C.C. and Saier,M.H.,Jr (1996) Structural and phylogenetic analysis of the MotA and MotB families of bacterial flagellar motor proteins. *Res. Microbiol.*, **147**, 317–332.
24. Macfadyen,L.P., Dorocicz,I.R., Reizer,J., Saier,M.H.,Jr and Redfield,R.J. (1996) Regulation of competence development and sugar utilization in *Haemophilus influenzae* Rd by a phosphoenolpyruvate:fructose phosphotransferase system. *Mol. Microbiol.*, **21**, 941–952.
25. Dinh,T., Paulsen,I.T. and Saier,M.H.,Jr (1994) A family of extracytoplasmic proteins that allow transport of large molecules across the outer membranes of Gram-negative bacteria. *J. Bacteriol.*, **176**, 3825–3831.
26. Paulsen,I.T., Beness,A.M. and Saier,M.H.,Jr (1997) Computer-based analyses of the protein constituents of transport systems catalysing export of complex carbohydrates in bacteria. *Microbiology*, **143**, 2685–2699.
27. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
28. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
29. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.