Review

Journal of Molecular Microbiology and Biotechnology

J Mol Microbiol Biotechnol 2006;11:291–301
DOI: 10.1159/000095631

Phylogenetic Analysis of General Bacterial Porins: A Phylogenomic Case Study

Thai X. Nguyen Eric R. Alegre Scott T. Kelley

Department of Biology, San Diego State University, San Diego, Calif., USA

Key Words
Bioinformatics · Outer membrane protein · *ompF* · *ompC* · Phylogeny · Porin

Abstract
Bacterial porin proteins allow for the selective movement of hydrophilic solutes through the outer membrane of Gram-negative bacteria. The purpose of this study was to clarify the evolutionary relationships among the Type 1 general bacterial porins (GBPs), a porin protein subfamily that includes outer membrane proteins *ompC* and *ompF* among others. Specifically, we investigated the potential utility of phylogenetic analysis for refining poorly annotated or misannotated protein sequences in databases, and for characterizing new functionally distinct groups of porin proteins. Preliminary phylogenetic analysis of sequences obtained from GenBank indicated that many of these sequences were incompletely or even incorrectly annotated. Using a well-curated set of porins classified via comparative genomics, we applied recently developed bayesian phylogenetic methods for protein sequence analysis to determine the relationships among the Type 1 GBPs. Our analysis found that the major GBP classes *(ompC*, *phoE*, *nmpC* and *ompN)* formed strongly supported monophyletic groups, with the exception of *ompF*, which split into two distinct clades. The relationships of the GBP groups to one another had less statistical support, except for the relationships of *ompC* and *ompN* sequences, which were strongly supported as sister groups. A phylogenetic analysis comparing the relationships of the GenBank GBP sequences to the correctly annotated set of GBPs identified a large number of previously unclassified and mis-annotated GBPs. Given these promising results, we developed a tree-parsing algorithm for automated phylogenetic annotation and tested it with GenBank sequences. Our algorithm was able to automatically classify 30 unidentified and 15 misannotated GBPs out of 78 sequences. Altogether, our results support the potential for phylogenomics to increase the accuracy of sequence annotations.

Introduction

Gram-negative bacteria are distinguishable from Gram-positive bacteria by the presence of an outer membrane. This membrane serves as a selective permeation barrier that restricts the movement of hydrophilic solutes in and out of the cell [Koebnik et al., 2000; Nikaido, 2003]. The movement of solutes across the membrane is made possible by channel-forming proteins. The general bacterial porins (GBPs) comprise one such class of channel-forming proteins found in members of the gamma-proteobacteria, such as *Escherichia coli*, *Shigella*, *Salmonella*, *Yersinia* and others [Koebnik et al., 2000; Schulz, 2002]. These non-specific permeation porins are the most abundant outer membrane proteins of enteropathogenic bacteria [Blasband et al., 1986; Blasband and Schnaitman,

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1464–1801/06/0116–0291$23.50/0

Accessible online at:
www.karger.com/mmb

Scott T. Kelley
Department of Biology, San Diego State University
San Diego, CA 92182-4614 (USA)
Tel. +1 619 594 5371, Fax +1 619 594 4767
E-Mail skelley@sciences.sdsu.edu

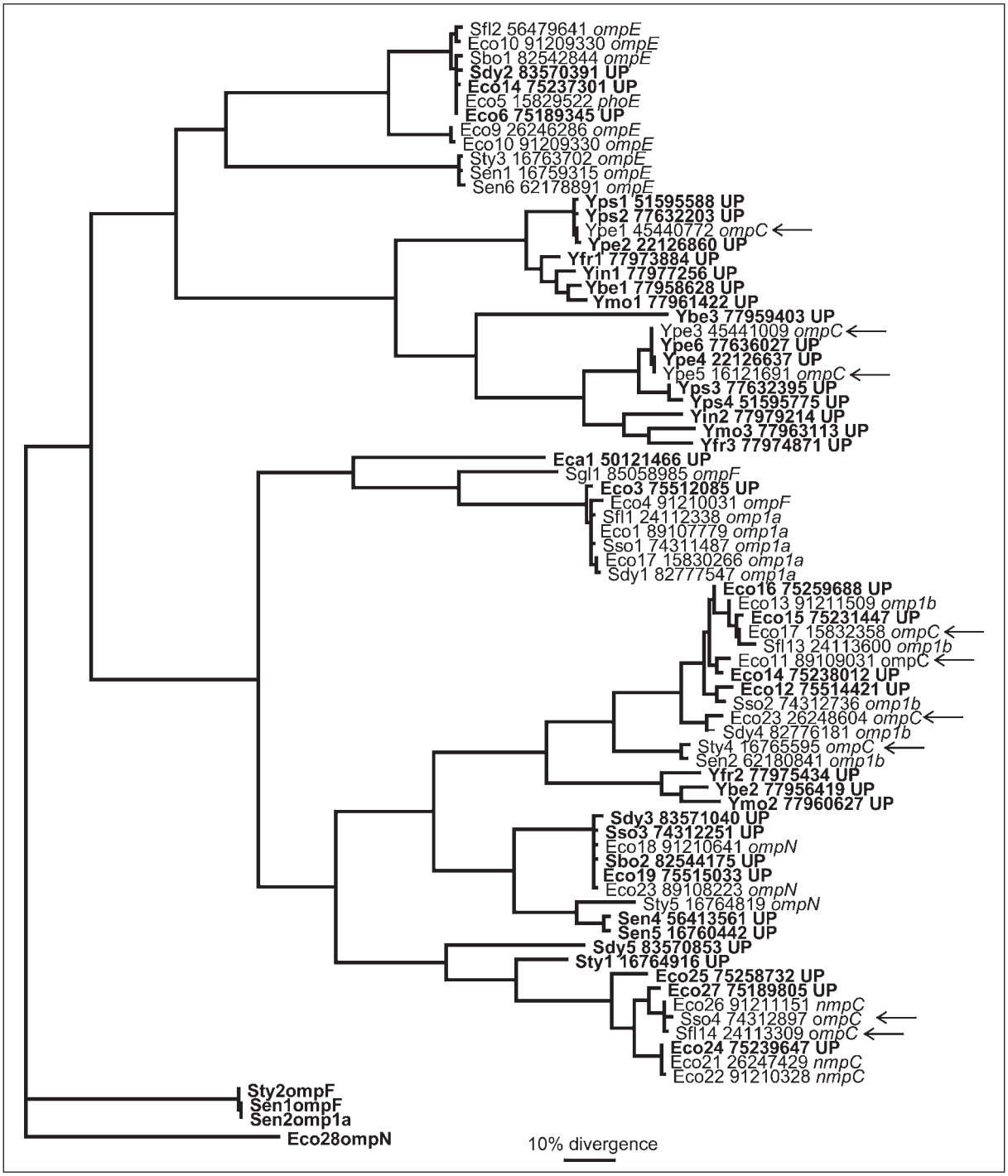

Fig. 1. Preliminary NJ analysis of putative Type 1 GBPs obtained from a BLAST search of GenBank using the *E. coli ompF* sequence as the query. The main purpose of the figure is to show the considerable discrepancies between GenBank annotations and the phylogeny of the sequences. Names in boldface indicate protein sequences identified only as porin-like (UP = unknown porin), while arrows highlight sequences annotated as *ompC* porins. The sequence information at the tips of the branches includes a three-letter code for the bacterial species (see below) with a unique arbitrary identifying number for comparing trees in figures 3 and 4. The name also includes the GenBank Identifier (GI) for the sequence, and gene annotation information provided in the GenBank file. Eca = *Erwinia carotovora*; Eco = *E. coli*; Sbo = *Shigella boydii*; Sdy = *Shigella dysenteriae*; Sen = *Salmonella enterica*; Sfl = *Shigella flexneri*; Sgl = *Sodalis glossinidius*; Sso = *Shigella sonnei*; Sty = *Salmonella typhimurium*; Ybe = *Yersinia bercovieri*; Yfr = *Yersinia frederiksenii*; Ymo = *Yersinia mollaretii*; Ype = *Yersinia pestis*; Yps = *Yersinia pseudotuberculosis*.

1987; Schulz, 2002]. The monomeric porin proteins form a stable trimeric channel that allows passive diffusion of nutrients across the outer membrane, and this trimer can also facilitate adhesion, invasion, and parasitism of pathogenic bacteria [Williams et al., 2000].

The three best-studied GBPs in *E. coli* include *ompF*, *ompC*, and *phoE*, and they differ from one another in their solute selectivity [Nikaido, 2003]. The expression of these well-characterized porins is affected by osmolarity, temperature, available carbon sources, and phosphate concentration, and these conditions have been used to characterize other porins, such as *nmpC*, and the LC porins [Nikaido, 2003]. The LC porin and *nmpC* genes are located on lambdoid bacteriophage and defective lambdoid prophage that have been integrated into bacterial genomes [Blasband et al., 1986; Blasband and Schnaitman, 1987; Prilipov et al., 1998]. The *ompF*, *ompC*, *phoE*, *nmpC*, and LC porins have all been classified as Type 1 GBPs, according to the Transport Classification Database (TCDB) [Saier et al., 2006], and we adopt this classification throughout this paper.

The purpose of this study was to determine the phylogenetic relationships among the Type 1 GBPs in order to better understand their evolution and ultimately assist the development of a broad-spectrum GBP vaccine antigen [Singh et al., 1995]. However, preliminary phylogenetic analysis of GBP-like porins obtained from a GenBank BLAST search indicated that a substantial number of protein sequences identified as *ompF*, *ompC*, or other types of general class porins were either poorly annotated or mis-annotated (fig. 1). Most of the bacterial sequences we procured from GenBank had presumably been annotated using the BLAST algorithm [Altschul et al., 1990]. The BLAST algorithm is arguably the most powerful and useful tool in bioinformatics, and has been used to functionally annotate millions of genes saving untold hours of experimentation and providing remarkable insight into biological systems. Although this algorithm is both deceptively simple and remarkably powerful, researchers have recognized that the BLAST algorithm cannot reliably distinguish between orthologous (sequences related through common ancestry) and paralogous (sequence similarity due to an ancestral duplication event) genes [Barbazuk et al., 2000; Chiu et al., 2006; Daubin et al., 2002; Srinivasan et al., 2005]. Determination of orthology or paralogy is critically important because paralogous genes often have distinct functional roles in organisms (e.g., *ompF*, *ompC*). Phylogenetic analyses, on the other hand, easily distinguish orthologs from paralogs given sufficient sampling of related sequences, and these methods have often been used to characterize new functional groups of proteins [Barbazuk et al., 2000; Kelley and Thackray, 1999; Yi et al., 1999].

Given the utility of phylogenetic analyses for classifying orthologs and paralogs, we set forth to determine the effectiveness of newly developed phylogenetic methods for establishing the relationships among the Type 1 GBPs. Using a set of GBPs that had been annotated using a combination of BLAST similarity and comparative genomic position analysis, we first determined whether phylogenetic approaches could accurately recover known groupings with high confidence. In other words, did the correctly annotated *ompC*, *ompF*, and other GBPs form strongly supported monophyletic groups? Second, we asked whether phylogenetic methods could determine the GBP group affiliation of unidentified porin-like sequences and also correct erroneous annotations. Finally, using newly developed bayesian phylogenetic methods that incorporate advanced models of protein evolution [Ronquist and Huelsenbeck, 2003], we investigated the evolutionary relationships among the various Type 1 GBP classes and attempted to detect new classes of uncharacterized GBPs. In the process of answering these questions, we also developed a phylogenetic algorithm for automatically annotating new sequences given a correctly annotated set of related sequences. We demonstrate the effectiveness of this 'phylogenomic' approach using porin-related protein sequences obtained from GenBank, and discuss its potential use in automated gene annotation. Our results suggest that automated phylogenetic methods, combined with BLAST methods and cross-genome comparisons, could be highly effective for improving the quality of gene functional annotations and reducing annotation error propagation in sequence databases.

Results

Multiple sequence alignments proved to be of high quality, with few insertions or deletions. Most of these insertions or deletions (indels) were in the variable extracellular regions of the porin, regions which are known to undergo relatively rapid evolutionary change [Nikaido, 2003]. For example, out of 486 amino acid alignment positions in the multiple sequence alignment used to estimate the phylogeny in figure 3, approximately 10% of the alignment positions contained gaps. Even in the regions with gaps, the alignment showed high regions of similar-

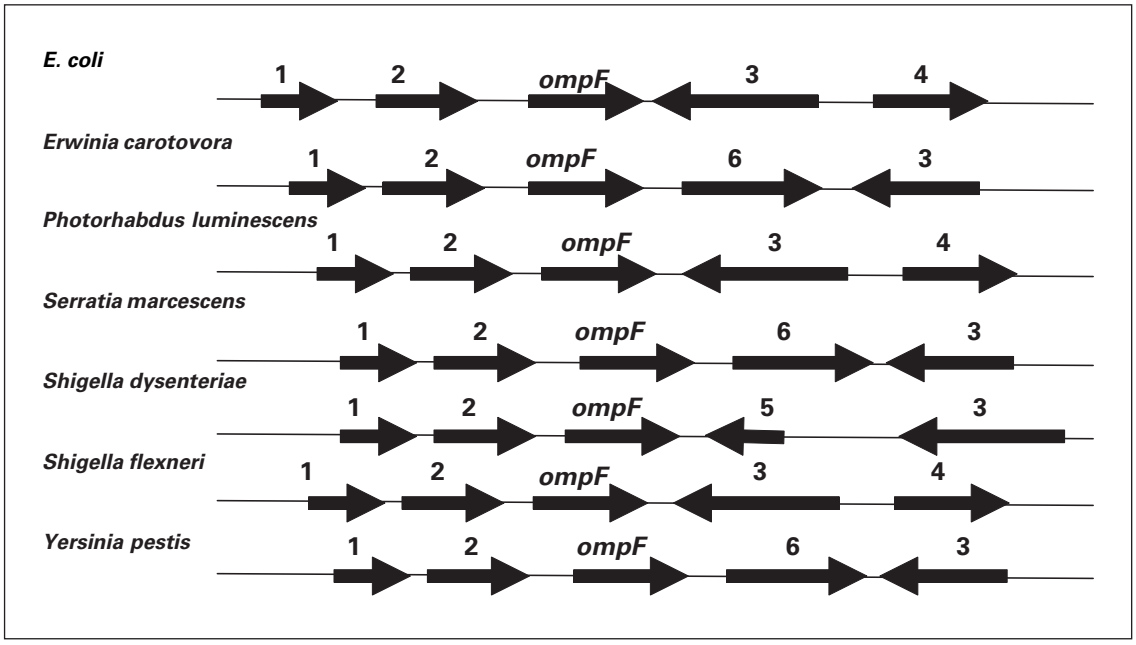

Fig. 2. Example analysis showing the relative genome position of *ompF*-like sequences in seven genomes using the SEED database. The arrows show the direction of predicted open reading frames (ORFs), and the numbers above the ORFs indicate prediction functions as follows: (1) uridine kinase *(udk)*; (2) deoxycytidine triphosphate deaminase *(dcd)*; (3) integral membrane protein/hemolysin *(yegH)*; (4) putative polysaccharide export protein *(wza)*; (5) helix-turn-helix motif *(ECs2870)*, and (6) anaerobic C4-dicarboxylate transporter *(dcuC)*.

ity, and most of the large gaps were due to only a few sequences that had long insertions relative to all the others. We also found greater numbers of indels and a high level of dissimilarity when comparing the alignment of Type 1 GBPs to other GPB types (e.g., *ompU*, *ompP2*), confirming the appropriateness of these sequences as outgroups.

Figure 2 shows an example genome homology analysis using the *E. coli ompF* protein to identify homologous sequences in other genomes. The SEED database tools allowed us to identify probable orthologs across genomes for *ompF*, *phoE*, *nmpC*, LC and *ompC* (table 1). After collecting sequences from three databases (GenBank, TIGR and the SEED) we used a Neighbor-joining (NJ) analysis, and the underlying pairwise distance matrix, to remove redundant sequences and to select a set of porins representing the greatest diversity in terms of both sequence and genome diversity for phylogenetic analysis. These sequences are described in table 1.

Figure 3 shows the results of bayesian phylogenetic analysis of the sequences presented in table 1. A preliminary bayesian analysis (100,000 MCMC generations) using multiple amino acid substitution models, found that the Wag model [Whelan and Goldman, 2001] had the highest posterior probability (p = 0.703), and this model was used for the rest of the analyses. With the exception of *ompF*, all of the sequences identified as homologous based on genome position clustered together in highly supported monophyletic groups. The *phoE*, *ompC*, and *nmpC* sequences all formed monophyletic groups with posterior probabilities of 1.0, 1.0 and 0.79, respectively, and high MP and NJ bootstrap support (fig. 3). Closer analysis of the sequences also identified a strongly supported monophyletic group of sequences from 8 different genomes that included a sequence identical to the originally identified *ompN* sequence (fig. 3). The *ompF* sequences identified from the SEED, on the other hand, did not form a single monophyletic group. Rather, they appeared to be polyphyletic and included two, and perhaps three, separate clades all with high posterior probability support (fig. 3).

The phylogenetic analysis also provided some limited insight into the relationships among the porin clusters. We found reasonable levels of support for a sister-group relationship between *ompC* and *ompN* (posterior probability of 0.88; fig. 3), and this relationship was bolstered by the addition of the GenBank sequences (fig. 4). Fig-

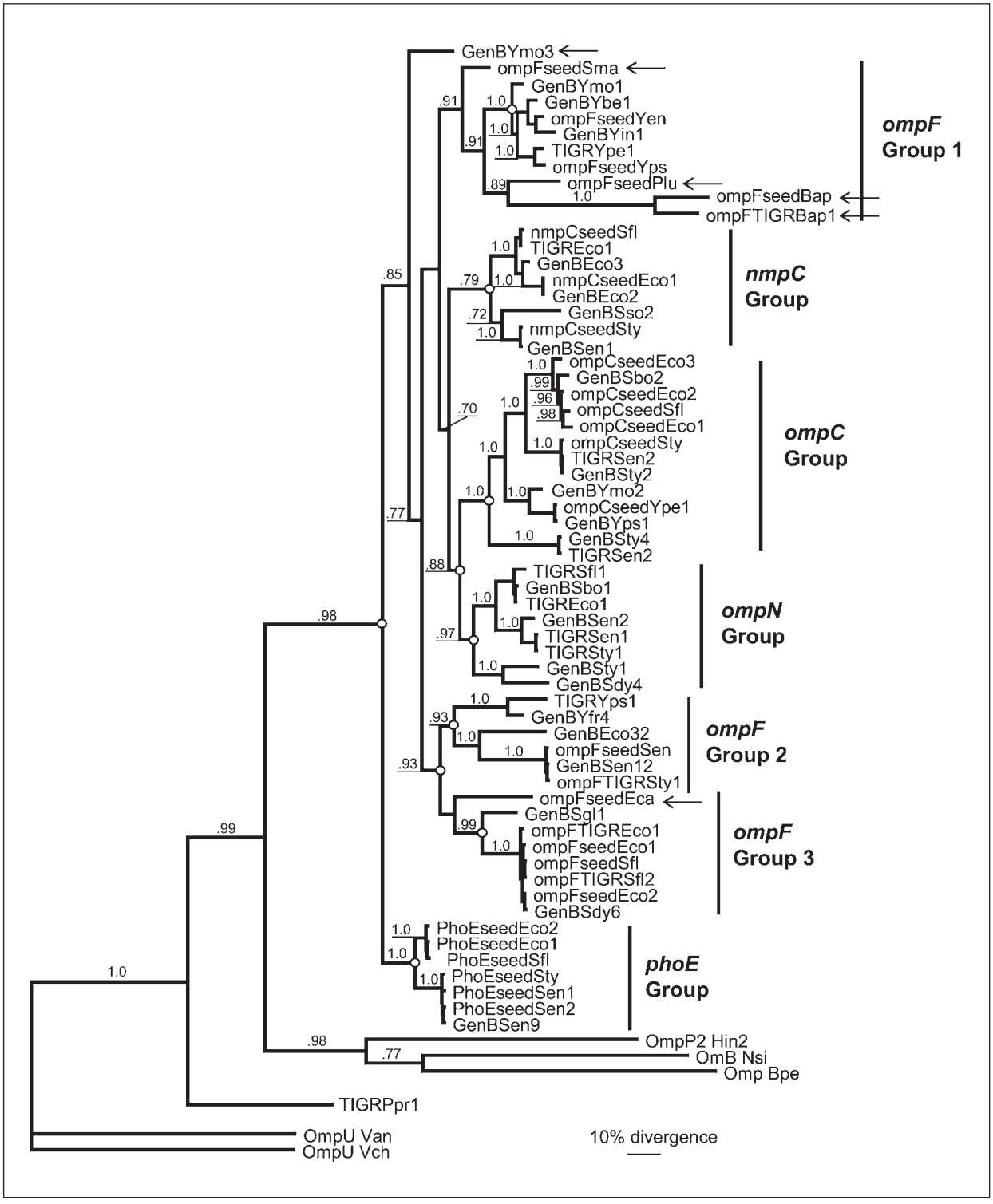

Fig. 3. Bayesian phylogenetic analysis of Type 1 GBPs identified from three databases (table 1). The sequences selected for this analysis were selected to maximize both sequence and bacterial species diversity. Code names that begin with a gene name and include the word 'seed' (e.g., ompCseedYps) had been annotated using the SEED genome comparison tools (see fig. 2). The values indicate posterior probabilities for particular nodes, with 1.0 being the maximum probability. Circles show nodes with MP and NJ bootstrap support exceeding 70%. Arrows indicate sequences that had different relationships to the rest of the sequences in the trees produced using MP or NJ methods.

Table 1. Information on protein sequences used in phylogenetic analyses shown in figure 3 organized by source database

Sequence ID	Annotation	Organism	Length	GI
TIGR database				
TIGREco1	*nmpC*	*Escherichia coli* CFT073	342	26108604
TIGREco1	*ompN*	*Escherichia coli* CFT073	377	26108086
TIGRSen1	*ompN*	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis	377	62127693
TIGRSty1	*ompN*	*Salmonella typhimurium* LT2	377	16419993
TIGRSfl1	*omp1b*	*Shigella flexneri* 2a str. 301	298	24052161
TIGRYpe1	Unknown porin	*Yersinia pestis* KIM	376	21959649
TIGRPpr1	Unknown porin	*Photobacterium profundum* SS9	339	46916736
TIGRSen2	*ompC*	*Salmonella enterica* subsp. *enterica* serovar Typhi	378	16503494
TIGRYps1	*ompC*	*Yersinia pseudotuberculosis*	360	51589572
TIGRBap1	*ompF*	*Buchnera aphidicola*	369	21623255
TIGREco1	*ompF*	*Escherichia coli* CFT073	362	26107356
TIGRSty1	*omp1a*	*Salmonella typhimurium* LT2	363	16419512
GenBank				
GenBEco2	Unknown porin	*Escherichia coli* K12	375	16128536
GenBEco3	*nmpC*	*Escherichia coli* CFT073	380	26247429
GenBSbo1	Unknown porin	*Shigella boydii* BS512	377	75178657
GenBSbo2	Unknown porin	*Shigella boydii* BS512	376	75176154
GenBSdy4	Unknown porin	*Shigella dysenteriae* 1012	395	83569514
GenBSdy6	*omp1a*	*Shigella dysenteriae* 1012	362	82777547
GenBSen1	Unknown porin	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis	362	62180142
GenBSen2	Unknown porin	*Salmonella enterica* subsp. *enterica* serovar Typhi	383	16760442
GenBSen9	*ompE*	*Salmonella enterica* subsp. *enterica* serovar Paratyphi	350	56414552
GenBSgl1	*ompF*	*Sodalis glossinidius*	368	85058985
GenBSso2	Unknown porin	*Shigella sonnei*	366	74312162
GenBSty1	Unknown porin	*Salmonella typhimurium* LT2	398	16765331
GenBSty2	*ompC*	*Salmonella typhimurium* LT2	378	16765595
GenBSty4	Unknown porin	*Salmonella typhimurium* LT2	372	16764875
GenBYbe1	Unknown porin	*Yersinia bercovieri* ATCC 43970	374	77956526
GenBYfr4	Unknown porin	*Yersinia frederiksenii* ATCC 33641	361	77975176
GenBYin1	Unknown porin	*Yersinia intermedia* ATCC 29909	376	77979214
GenBYmo1	Unknown porin	*Yersinia mollaretii* ATCC 43969	367	77963113
GenBYmo2	Unknown porin	*Yersinia mollaretii* ATCC 43969	372	77960627
GenBYmo3	Unknown porin	*Yersinia mollaretii* ATCC 43969	371	77961422
GenBYps1	*ompC*	*Yersinia pseudotuberculosis*	374	51595605
GenBSen12	*omp1a*	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis	365	62179526
SEED database				
nmpCseedSfl	LC porin	*Shigella flexneri* 2a str. 301	360	24113309
nmpCseedEco1	LC porin	*Escherichia coli* K12	375	1786765
nmpCseedSty	Unknown porin	*Salmonella typhimurium* LT2	362	16420094
phoEseedEco1	Unknown porin	*Escherichia coli* E24377A	353	75189345
phoEseedEco2	Unknown porin	*Escherichia coli* CFT073	353	26246286
phoEseedSen1	*phoE*	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis	350	62178891
phoEseedSen2	*phoE*	*Salmonella enterica* subsp. *enterica* serovar Typhi	350	29142912
phoEseedSty	*phoE*	*Salmonella typhimurium* LT2	350	16418821
phoEseedSfl	*phoE*	*Salmonella typhimurium* LT2	351	30061858
ompCseedEco1	*omp1b*	*Escherichia coli* K12	367	16130152
ompCseedSty	*ompC*	*Salmonella typhimurium* LT2	378	16765595
ompCseedEco2	*ompC*	*Escherichia coli* O157:H7	367	13362573
ompCseedSfl	*omp1b*	*Shigella flexneri* 2a str. 301	373	24113600
ompCseedEco3	*ompC*	*Escherichia coli* CFT073	375	26248604
ompCseedYpe1	*ompC*	*Yersinia pestis* CO92	374	16121511
ompFseedYen	Unknown porin	*Yersinia enterocolitica* 8081	374	N/A

Table 1 (continued)

Sequence ID	Annotation	Organism	Length	GI
ompFseedSma	Unknown porin	*Serratia marcescens* Db11	365	N/A
ompFseedBap	Unknown porin	*Photorhabdus asymbiotica* subsp. *asymbiotica*	366	N/A
ompFseedEco1	*omp1b*	*Escherichia coli* K12	362	16128896
ompFseedEco2	*omp1a*	*Escherichia coli* O157:H7	362	13360471
ompFseedEca	Unknown porin	*Erwinia carotovora* subsp. *atroseptica*	370	49611992
ompFseedPlu	*ompN*	*Photorhabdus luminescens* subsp. *laumondii* TTO1	388	37525686
ompFseedSen	*ompF*	*Salmonella enterica* subsp. *enterica* serovar Typhi	363	29142359
ompFseedSfl	*omp1a*	*Shigella flexneri* 2a str. 301	362	30062464
ompFseedYps	Unknown porin	*Yersinia pseudotuberculosis*	360	51595775

The annotations refer to information in GenBank files.
GI = GenBank identifier; N/A = not available.

ure 4 shows the results of a phylogenetic analysis showing the relationships of the sequences identified using the SEED database to the porin-like protein sequences used to make figure 1 that were obtained in our initial BLAST search of GenBank. The phylogeny indicated that many of these sequences were closely related to the SEED identified proteins (fig. 4). Given this phylogenetic analysis, and the SEED classifications, we 're-annotated' these proteins using our phylogenetic annotation algorithm, diagrammed in figure 5, and the phylogenetic annotation algorithm successfully identified the Type 1 GBP group membership of almost all the GenBank sequences from the tree in figure 4. Out of the 78 GenBank sequences, the algorithm identified the GBP group membership of 30 unidentified GBP-like protein sequences, corrected the annotation of 15 others (see the fig. 4 legend). Nine of the sequences did not belong within a known GBP phylogenetic group (Ybe3, Ymo1, Ype1, Yin1, Yfr1, Yps1, Yps2, Ype1, and Ype2; fig. 4), and the other 24 had been correctly annotated in GenBank, with the possible exception of the *ompE* annotations, which were originally considered to be functionally different from *phoE* [Chart et al., 1993].

Discussion

The results of our phylogenetic study showed the potential for phylogenomics to enhance our understanding of protein evolution and improve protein function prediction. The bayesian phylogenetic analysis found strong support for relationships within the Type 1 GBPs (fig. 3). All of the sequences identified by genomic position, with the exception of *ompF*, formed strongly supported monophyletic groups. These included the *ompC*, *phoE* and *nmpC* types of porins, as well as the *ompN* class of porins (fig. 3). These results did not change appreciably when we repeated the phylogenetic analysis excluding the ~10% alignment positions with large numbers of gaps (data not shown). The fact that the majority of the porins identified based on genome position (e.g., fig. 2) formed clearly identifiable monophyletic groups supports the notion that phylogenetic methods can accurately classify GBP orthologs.

The bayesian phylogenetic analysis also shed some light on the relative relationships among the Type 1 GBPs. For example, we found strong support for the sister-group relationship of *ompC* and *ompN* porins (fig. 3, 4), supporting the conclusions of Prilipov et al. [1998] that *ompN* and *ompC* were biochemically similar but still distinct types of GBPs. We did not initially include *ompN* as a distinct group of Type 1 GBPs because they had not been identified as such in the TCDB. However, we discovered that one of the sequences we obtained from the TIGR database was identical to the originally identified *ompN* sequence, and appeared to be part of a larger phylogenetic group of putative *ompN* sequences from other genomes (fig. 3). Altogether, we identified 7 other orthologous *ompN* sequences based on genome position in the SEED database, and these were used in the phylogenetic analysis shown in figure 4.

The *ompF* porins were the one major group that did not form a monophyletic group as expected based on the analysis of genome position. Instead, these proteins split

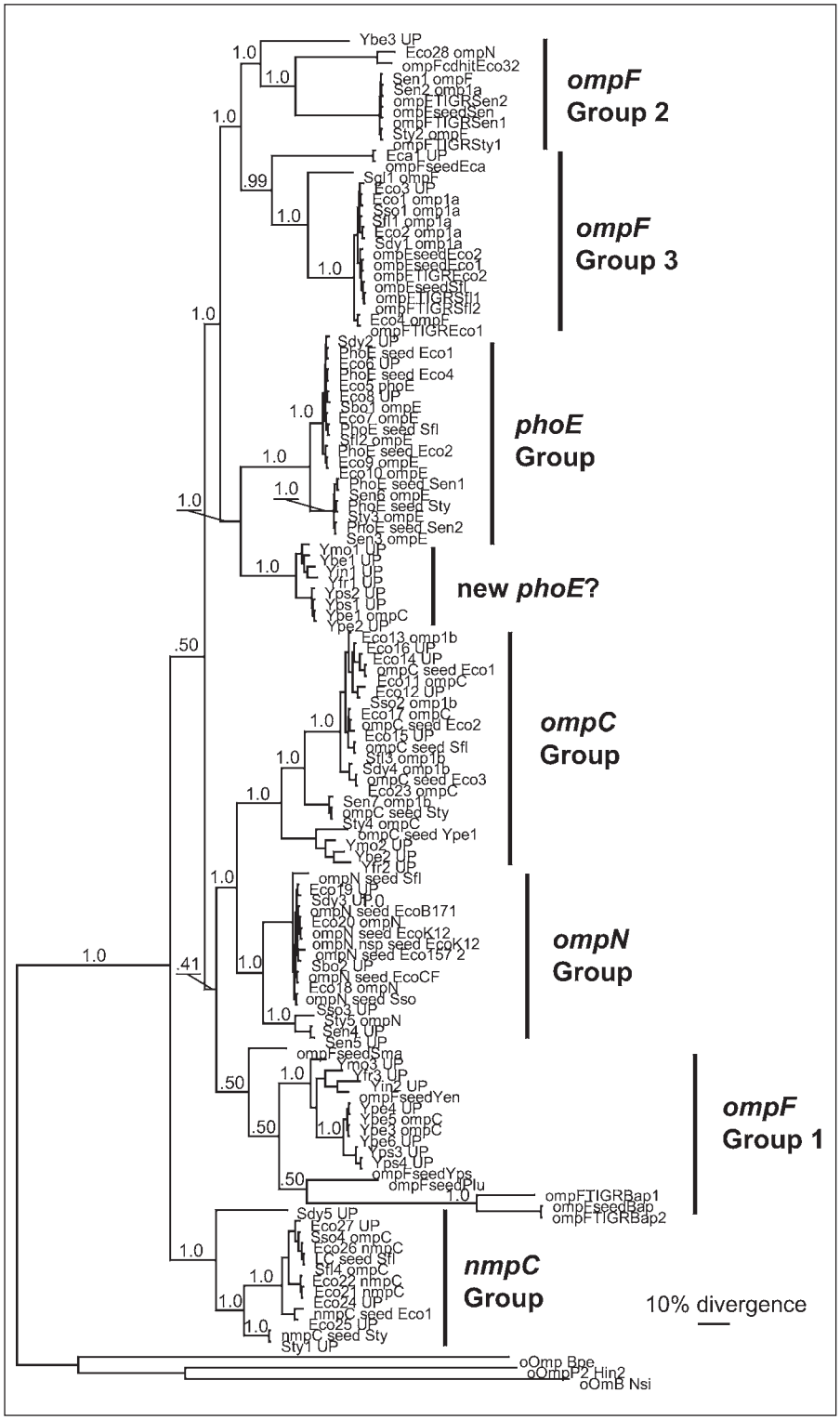

Fig. 4. Phylogenetic re-classification of Type 1 GBPs found in GenBank were used to make the tree in figure 1. The bayesian phylogenetic analysis compared the relationships of the figure 1 sequences to the SEED annotated sequences (table 1). This analysis also included five *ompN* sequences from the SEED database. The values indicate posterior probabilities for particular nodes, with 1.0 being the maximum probability. The vertical lines indicate various monophyletic GBP cluster. The algorithm detailed in figure 5 readily annotated all the GenBank sequences belonging to the indicated groups. For example, the unidentified GBPs Eco19, Sdy3, Sbo2, Sso3, Sen4 and Sen5 sequences were all annotated as *ompN* porins.

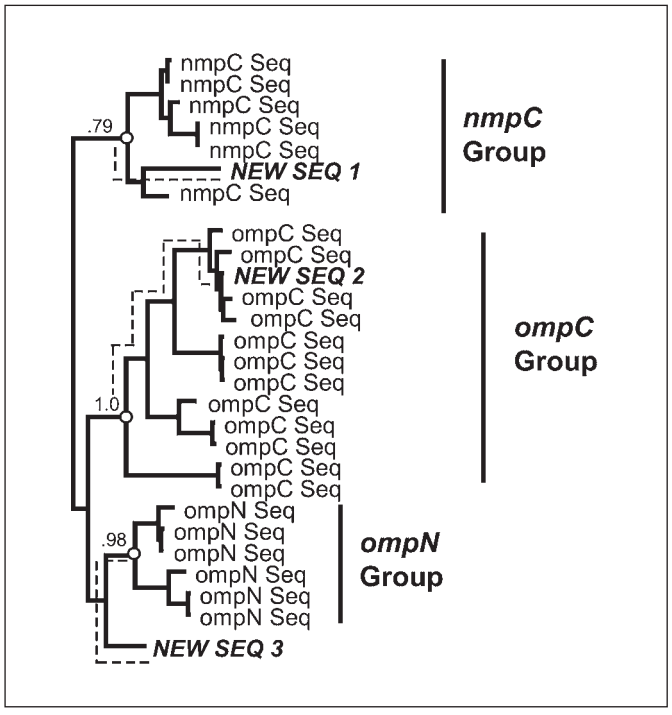

Fig. 5. Diagram of the phylogenomic annotation algorithm. The figure shows hypothetical phylogenetic relationships of three unidentified GBP-like proteins, labeled NEW SEQ1, NEW SEQ 2 and NEW SEQ 3, to a set of correctly annotated *nmpC*, *ompC* and *ompN* sequences. The dashed lines indicate the paths of the tracebacks that identified the position of the sequence in the tree relative to the internal nodes that contain all members of a particular group. The numbers at the nodes indicate the bayesian posterior probability for the three basal nodes. Using this algorithm, NEW SEQ 1 and NEW SEQ2 would be identified as members of the *nmpC* and *ompC* clades, respectively, while NEW SEQ 3 would be most closely affiliated with the *ompN* clade but could not be said to belong to that group.

into two distinct monophyletic groups, each of which had strong statistical support (fig. 3). The bayesian phylogenetic analysis indicated that these two groups were closely related and were only separated by one node on the phylogenetic tree. However, the *ompF* group appears to have a complex evolutionary history, and the *ompF* genes found in the same relative genomic position may have distinct biochemical properties. A follow-up analysis that added 78 more GBP sequences from GenBank also found the groups to be paraphyletic (fig. 4). These results suggest that, in combination with genomic information, phylogenetic analysis could be a potentially useful tool for identifying new protein functions even with heavily characterized proteins, such as *ompF*.

Once we had demonstrated the effectiveness of the phylogenetic approach for classifying porins, we then applied the same approach to check the annotation of the sequences in figure 1 collected from GenBank. As figure 4 shows, the phylogenetic approach readily classified GBP proteins given a set of known sequences, such as the ones from the SEED database. After adding the GenBank sequences, the major Type 1 GBP groups (aside from *ompF*) remained monophyletic (fig. 4), and many of the GenBank sequences were closely related to these groups. Using the strongly supported relationships shown in figure 4, we were able to: (1) classify a large number of sequences in the database that had been identified as 'unknown' porins, and (2) identify a number of apparent mistakes in the database. Interestingly, we found that the addition of the large numbers of GenBank sequences to the tree increased the statistical support at some deeper nodes of the tree, suggesting that more sampling might be helpful in resolving the relationship among the GBPs.

Given these promising results, we also developed a computer algorithm, diagrammed in figure 5, to automatically classify sequences based on their relationships to a set of known sequences (e.g., the porins identified by genomic position). If a sequence belonged to a highly supported monophyletic group of Type 1 GBPs, such as *ompC*, *phoE* or *ompN*, the tree-parsing algorithm successfully identified the group affiliations (fig. 4, 5). Out of the 78 GenBank sequences we analyzed, our phylogenetic algorithm identified the GBP group membership for 30 unidentified porin-like sequences and corrected the annotation of 15 other sequences (fig. 4). The number of mis-annotations might be higher depending on the status of the *ompE* annotated GBPs. The *ompE* porins were originally thought to be distinct from *phoE* porins [Chart et al., 1993], though our phylogeny suggests otherwise.

Our algorithm could not precisely identify nine of the other sequences because they did not belong within a monophyletic group of known GBPs. Eight of these were identified as 'new phoE?' in figure 4 because they comprise of a strongly supported sister group with the other *phoE* sequences, and the algorithm correctly identified the nearest GBP group for all of these sequences. The fact that 49% of the GenBank sequences we obtained for just the GBPs were only partially annotated, and 20% were likely mis-annotated, indicates that a high proportion of automated database annotations may be incomplete or erroneous. Similar database issues have been pointed out by a number of other authors who have lamented the state

of annotations in GenBank and other databases [Ouzounis and Karp, 2002]. Our results suggest that a phylogenomic approach may be especially helpful for resolving and correcting annotations and reducing problems of error propagation.

Future work on this topic will include development of an automated annotation system that uses multiple sequence alignments and phylogenetic analyses to refine functional annotations with porins and other proteins. We are currently testing the effectiveness of our algorithm with more bacterial sequences, and we need to compare the effectiveness of our methods with other recently developed phylogenomic approaches [Chiu et al., 2006; Srinivasan et al., 2005]. Nonetheless, our preliminary study of GBPs suggests that implementing an automated phylogenomic approach, combined with genomic-position analyses and BLAST searches, could significantly enhance the accuracy of protein sequence annotations.

Experimental Procedures

Sequence Collection, Identification and Multiple Sequence Alignment

Amino acid sequences were obtained from three databases: NCBI (GenBank), TIGR (http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi) and the SEED (http://theseed.uchicago.edu/FIG/index.cgi). We used BLAST searches with known *E. coli* GBP proteins to identify porin sequences in GenBank and TIGR (table 1). We used the 'Pins' function in the SEED database to identify the relative genomic position of putative *ompF*, *phoE*, *nmpC*, LC and *ompC* homologs in genomes available in that particular database.

We also identified five potential outgroup sequences using the TCDB database (http://www.tcdb.org/tcdb/superfamily.php). According to TCDB, the GBPs comprises nine distinct groups based on biochemical and sequence properties (see the GBP section of the TCDB: http://www.tcdb.org/tcdb/index.php?tc=1.B.6). Since *ompF*, *phoE*, *nmpC*, LC and *ompC* all belong to the first biochemical cluster, we selected sequences from the second, third and fourth clusters as outgroup sequences: *ompU* of *Vibrio cholerae* (GI: 12644367), *ompU* of *Listonella (Vibrio) anguillarum* (GI: 75446970), *ompP2* of *Haemophilus influenzae* (GI: 3914220), *omp* porin of *Bordetella pertussis* (GI: 1709465), and the *por* protein from *Neisseria sicca* (GI: 266700). Protein sequences were aligned using clustalW [Chenna et al., 2003] and inspected manually to insure high quality.

Phylogenetic Analyses

We used MrBayes version 3.1 to perform bayesian phylogenetic analyses [Ronquist and Huelsenbeck, 2003]. MrBayes provides a comprehensive set of protein evolution models and the ability to estimate the model that best fits a given dataset. To determine the highest likelihood model of protein evolution for our data, we ran the MCMC sampler for 100,000 generations using the mixed amino acid model. After determining the best-fit protein model, we ran the MCMC sampler for 3 million generations using the fixed model.

We used the PAUP* program [Swofford, 1998] to perform Maximum Parsimony (MP), NJ and bootstrap analyses. Shortest MP trees were found using a heuristic search strategy using TBR (Tree Bisection-Reconstruction) branch swapping. One hundred random addition sequence heuristic replicates were performed to find the shortest tree for each data set. The bootstrap analyses were performed under both MP and NJ criterion. For the MP bootstrap analysis, we ran 100 bootstrap replicates with 10 random addition heuristic searches performed per replicate (TBR branch-swapping). One thousand bootstrap replicates were performed under the NJ criterion. MP, NJ and bayesian trees were viewed and converted for graphical manipulation with TreeView 1.6.6 [Page, 1996].

Acknowledgements

We thank Roger Sabbadini, Kathleen McGuire and Stanley Maloy for encouragement and helpful suggestions on the course of study. We also thank Rob Edwards for assistance using the SEED database, Sujata Sovani for technical assistance, and Dean Ellis for comments on the manuscript. This work was supported by a grant from the California State University Program for Education and Research in Biotechnology (CSUPERB).

References

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.

Barbazuk, W.B., Korf, I., Kadavi, C., Heyen, J., Tate, S., Wun, E., Bedell, J.A., McPherson, J.D., and Johnson, S.L. 2000. The syntenic relationship of the zebrafish and human genomes. Genome Res 10:1351–1358.

Blasband, A.J., Marcotte, W.R., Jr., and Schnaitman, C.A. 1986. Structure of the lc and nmpC outer membrane porin protein genes of lambdoid bacteriophage. J Biol Chem 261: 12723–12732.

Blasband, A.J., and Schnaitman, C.A. 1987. Regulation in *Escherichia coli* of the porin protein gene encoded by lambdoid bacteriophages. J Bacteriol 169:2171–2176.

Chart, H., Frost, J.A., and Rowe, B. 1993. Expression of a new porin 'OmpE' by strains of *Salmonella enteritidis*. FEMS Microbiol Lett 109:185–187.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. 2003. Multiple sequence alignment with the Clustal series of programs. Nucl Acids Res 31:3497–3500.

Chiu, J.C., Lee, E.K., Egan, M.G., Sarkar, I.N., Coruzzi, G.M., and DeSalle, R. 2006. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics 22:699–707.

Daubin, V., Gouy, M., and Perriere, G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res 12:1080–1090.

Kelley, S.T., and Thackray, V.G. 1999. Phylogenetic analyses reveal ancient duplication of estrogen receptor isoforms. J Mol Evol 49: 609–614.

Koebnik, R., Locher, K.P., and Van Gelder, P. 2000. Structure and function of bacterial outer membrane proteins: barrels in a nutshell. Mol Microbiol 37:239–253.

Nikaido, H. 2003. Molecular basis of bacterial outer membrane permeability revisited. Microbiol Mol Biol Rev 67:593–656.

Ouzounis, C.A., and Karp, P.D. 2002. The past, present and future of genome-wide re-annotation. Genome Biol 3:Comment 2001.

Page, R.D. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12:357–358.

Prilipov, A., Phale, P.S., Koebnik, R., Widmer, C., and Rosenbusch, J.P. 1998. Identification and characterization of two quiescent porin genes, *nmpC* and *ompN*, in *Escherichia coli* BE. J Bacteriol 180:3388–3392.

Ronquist, F., and Huelsenbeck, J.P. 2003. MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.

Saier, M.H., Jr., Tran C.V., and Barabote, R.D. 2006. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucl Acids Res 34:D181–D186.

Schulz, G.E. 2002. The structure of bacterial outer membrane proteins. Biochim Biophys Acta 1565:308–317.

Singh, S.P., Singh, S.R., Williams, YU., Jones L., and Abdullah, T. 1995. Antigenic determinants of the *ompC* porin from *Salmonella typhimurium*. Infect Immun 63:4600–4605.

Srinivasan, B.S., Caberoy, N.B., Suen G., Taylor, R.G., Shah, R., Tengra F., Goldman, B.S., Garza, A.G., and Welch, R.D. 2005. Functional genome annotation through phylogenomic mapping. Nat Biotechnol 23:691–698.

Swofford, D. 1998. PAUP*: phylogenetic analysis using parsimony (* and other methods). Sinauer Associates, Sunderland, Mass.

Whelan, S., and Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–699.

Williams, K.M., Bigley, E.C., 3rd, and Raybourne, R.B. 2000. Identification of murine B-cell and T-cell epitopes of *Escherichia coli* outer membrane protein F with synthetic polypeptides. Infect Immun 68:2535–2545.

Yi, C.H., Terrett, J.A., Li, Q.Y., Ellington, K., Packham, E.A., Armstrong-Buisseret, L., McClure, P., Slingsby, T., and Brook, J.D. 1999. Identification, mapping, and phylogenomic analysis of four new human members of the T-box gene family: EOMES, TBX6, TBX18, and TBX19. Genomics 55:10–20.