Correspondence of Function and Phylogeny of ABC Proteins Based on an Automated Analysis of 20 Model Protein Data Sets

Georg Fuellen1,2,2* Michael Spitzer1,2,2 Paul Cullen4 and Stefan Lorkowski4,5
1Department of Medicine, AG Bioinformatics, University of Münster, Münster, Germany
2Division of Bioinformatics, Biology Department, University of Münster, Münster, Germany
3Integrated Functional Genomics, Interdisciplinary Center for Clinical Research, Münster, Germany
4Leibniz Institute of Arteriosclerosis Research, University of Münster, Münster, Germany
5Institute of Biochemistry, University of Münster, Münster, Germany

ABSTRACT Using our BLAST-based procedure RiPE (Retrieval-induced Phylogeny Environment), which automates the evolutionary analysis of a protein family, we assembled a set of 1138 ABC protein components [adenosine triphosphate (ATP)-binding cassette and transmembrane domain] from the protein data sets of 20 model organisms and subjected them to phylogenetic and functional analysis. For maximum speed, we based the alignment directly on a homology search with a profile of all known human ABC proteins and used neighbor-joining tree estimation. All but 11 sequences from Homo sapiens, Arabidopsis thaliana, Drosophila melanogaster, and Saccharomyces cerevisiae were placed into the correct subtree/subfamily, reproducing published classifications of the individual organisms. By following a simple “function transfer rule”, our comparative phylogenetic analysis successfully predicted the known function of human ABC proteins in 19 of 22 cases. Three functional predictions did not correspond, and 10 were novel. Predictions based on BLAST alone were inferior in five cases and superior in two. Bacterial sequences were placed close to the root of most subtrees. This placement coincides with domain architecture, suggesting an early diversification of the ABC family before the kingdoms split apart. Our approach can, in principle, be used to annotate any protein family of any organism included in the study. Proteins 2005;61:888–899.

© 2005 Wiley-Liss, Inc.

Key words: ABC proteins; phylogeny; evolution; homology search; annotation

INTRODUCTION

Predicting protein function from sequence (i.e., annotating a protein data set) is a very common task in bioinformatics.1 The basic idea is that homology implies functional analogy, transferring the function of characterized related proteins to the one under investigation.2 One such approach is based on bidirectional BLAST best hits3,4; this approach received some criticism that it may be misled by superficial similarity.5,6 Clustering of putative orthologs with assumed functional invariance is an alternative approach,7–11 as is pattern-based classification using motif libraries.12,13 Other methods rely on phylogenetic trees even though these are laborious to compute.14–21 The approach presented here is also based on phylogeny. However, we exploit both the radius with which an alignment can be generated directly from the results of a profile database search, and the speed of newer phylogeny software22 to allow us to analyze a very large number of sequences. If the genesis of the sequences follows a birth–death process, as is the case for many genes, there is good reason that increasing the number of sequences will increase phylogenetic accuracy.23

The phylogenetic analysis of protein sequences is complicated by the diversity of evolutionary processes that are acting on proteins, yielding problems such as unequal substitution rates, gene conversion, functional convergence, (tandem) duplications, and a mosaiclike subdivision of many proteins into distinct regions or domains.24,25 Regarding the latter aspect, a standard approach is the phylogenetic analysis of individual domains.26 However, in many cases, these domains are too short to obtain a reliable estimation of their evolutionary history. In the case of adenosine triphosphate (ATP)-binding cassette (ABC) proteins,27 the ABC domain included in the PRODOM domain database28 is divided into three parts, and for each part a different phylogenetic history is reported. Additional domains identified in ABC protein subfamilies also give rise to a tree of their own. It is unlikely that these different trees, each based on 50 residues or fewer, reflect evolutionary history. Efforts have been made to counter this kind of domain fragmentation, for example, the approach taken by the DOMO domain database.29 However, these larger domains are still too short and often lack specific characteristics of protein subfamilies. For example, in DOMO, the ABC protein domain lacks the

*Correspondence to: Georg Fuellen, Department of Medicine, c/o Division of Bioinformatics, Schlossplatz 4, 48149 Münster, Germany. E-mail: fuellen@uni-muenster.de

This article contains Supplementary Material, which can be found at www.interscience.wiley.com/jpages/0887-3585/suppmat/

© 2005 WILEY-LISS, INC.
C-terminal region that includes residues typical of the ABC subfamily G (see Table I for a list of ABC subfamily designations).\(^a\) Our approach was therefore to sample as much sequence as possible that is relevant for elucidating evolution and function, while excluding multiple copies of the same region in a single sequence, since these may distort phylogenetic analysis.

Our BLAST-based sampling and analysis procedure RiPE (Retrieval-induced Phylogeny Environment) reflects this aim.\(^b\) RiPE uses a database search to sample only the homologous parts of those sequences that display a sufficient chance of homology to the query. More specifically, we use a query profile (also called jumpstart alignment) consisting of the human ABC proteins to conduct a PSI-BLAST search.\(^c\) Many ABC proteins consist of two "halves" (i.e., a tandemly repeated domain arrangement). The so-called full-transporters consist of two similar ABC cassettes (symbols a and o) and two less similar transmembrane regions (symbols t and τ), mostly in the order "ta τα." Thus, "ταα" is a short form of the standard notation TMD\(_1\)ABC\(_1\)TMD\(_2\)ABC\(_2\), where ABC denotes the ATP-binding cassette and TMD the transmembrane domain. Such tandemly repeated domain arrangements are detected and give rise to separate entries in the query profile, and to separate database hits. Using RiPE, the choice of the query profile is determined by the question under investigation, allowing estimation of the evolutionary history of, for example, the whole human ABC protein family in the context of other model organisms or the evolution of a specific subfamily. In the former case, a query profile of all human ABC proteins was used whereby the five known human ABC subfamily G proteins were sufficient to start a homology search yielding all the information on which to base the phylogenetic and functional analysis of this subfamily.\(^d\) In both cases, the sequences from other organisms can then be classified with respect to the human system using only sequence parts that have homologous regions in human sequences. As a reference classification, we used Dassa and Bouige,\(^e\) which is not restricted to bacterial ABC proteins. As noted by Tomii and Kanehisa,\(^f\) there is a general consistency between different phylogeny-based classification schemes for bacterial ABC proteins, including their own, an earlier one by Saurin and Dassa,\(^g\) and the one by Tam and Saier.\(^h\) Many of the classes described in these schemes are associated with a certain functionality, and functional predictions have indeed been made by Tomii and Kanehisa for the case of bacterial sequences.\(^i\) However, these always refer to some general functionality of a certain class and are not as specific as our predictions based on the phylogenetic tree derived from the sequence information sampled using RiPE and the function transfer rule that we have developed. Using our approach, we were able to provide functional predictions for several previously uncharacterized proteins that were superior to predictions derived directly from BLAST searches. Moreover, the overall structure of our phylogenetic tree supports the notion that the ABC family diversified before the bacteria/archaea/eukarya split apart.

### MATERIALS AND METHODS

#### Retrieval of Sequence Databases

We retrieved the protein data sets from all eukarya for which nearly complete protein data sets were available in the summer of 2003 (see Table SIII in Supplementary Material). The selection of bacteria and archaea was based on the availability of published ABC inventories in case of *Escherichia coli* and *Mycobacterium tuberculosis*\(^i\),\(^j\) presumed close relatives of the endosymbionts that are believed to be the ancestors of mitochondria (*Rickettsia prowazekii*\(^k\),\(^l\) and chloroplasts (*Synechocystis* sp.\(^m\)) and phyletic diversity in case of the others.

#### Protein Family Sequence Profile Generation

We retrieved the sequences of all known human ABC proteins from the National Center for Biotechnology Information (NCBI) GenBank\(^n\) using the sequence accession codes listed on the Human ABC Transporter website (http://nutrigene.4t.com/humanabc.htm). As described in the Introduction, many ABC proteins feature tandem repeats of the ATP-binding cassette (a, α) and the transmembrane domain (t, τ). The presence of a second copy (τ or α) of a repeated domain (or the second domain arrangement "ταα" or "ατα") prevents achievement of consistent overall alignment since sequences with just one copy may align to either one of the two copies in the aα, ταα, or ατα arrangement. For this reason, we deleted the second copies...
<table>
<thead>
<tr>
<th>Human proteins</th>
<th>Function/substances transported</th>
<th>Nonvertebrate proteins of known function</th>
<th>Function/substances transported</th>
<th>Correspondence/prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(human protein)</td>
<td>(nonvertebrate protein)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Subfamily A (first halves)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Subfamily A (second halves)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1,2,3,4,7,12</td>
<td>● Phospholipids</td>
<td>● drrA (M. tuberculosis)</td>
<td>● Phthiocerol dimycocerosate (complex lipid), doxorubicin</td>
<td>Partial correspondence</td>
</tr>
<tr>
<td></td>
<td>● Estramustine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>● Lung surfactant pre-processing</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
|               | ● N-retinylidene-phosphatidyl-
|               | ethanolamine                     |                                        |                               |                          |
| A8,5,6,9,10   | ● Xenobiotics (e.g., estradiol-
|               | glucuronide, taurocholate (bile salt), leucotriene C4 (complex lipid)) | ● See A1,2,3,4 (see note)       | ● See A1,2,3,4           | Partial correspondence |
| Note: The ced-7 protein of C. elegans is found at the root of the tree of A8,5,6,9,10, but its presumed involvement in phospholipid movement has not been shown by experiment.52 |
| **Subfamily B (first halves/second halves in case of B1,4,5,11)** |
| B1,4,5        | ● Hydrophobic compounds, steroids, etc., phosphatidylcholine, cholines | ● mdr49 (D. melanogaster)           | ● Colchicines                | Multiple correspondences, 1/3 correct (i.e., colchicines) |
|               |                                 | ● mdr65 (D. melanogaster)             | ● α-amanitin                  |                          |
|               |                                 | ● ppg-1 (C. elegans)                 | ● Rhodamine 123               |                          |
| B11           | ● Bile salts, paclitaxol         | ● See B1,4,5                         | ● See B1,4,5                 | No correspondence       |
| B2,3,9        | ● Peptides                      | ● See B8,10                          | ● See B8,10                  | Multiple correspondences, 1/3 correct (i.e., peptides) |
|               |                                 | ● PMD1 (S. pombe)                    | ● Leptomycin B                |                          |
|               |                                 | ● MDL1 (S. cerevisiae)               | ● Peptides                    |                          |
| B6            | ● Iron metabolism               | ● See B7                             | ● See B7                     | Partial correspondence (see note) |
| B7            | ● Fe/S cluster metabolism       | ● ATM3 (A. thaliana)                 | ● Fe/S proteins              | Correspondence          |
| B8,10         | ● ?                             | ● MSEA (E. coli)                     | ● Lipid A, phospholipids, lipopolysaccharides | Prediction |
|               |                                 | ● MDR1 (A. thaliana)                 | ● Auxin (tryptophan derivative) |                          |
| Note: Involvement in iron metabolism is considered to correspond weakly to involvement in Fe/S cluster metabolism. |
| **Subfamily C (first halves)** |
| C1,2,3        | ● Sulfite, glutathione and glucuronide-conjugates of organic arsenite | ● MRP3,4,5 (A. thaliana)            | ● GS-conjugates, chlorophyll catabolites, glucuronides | Correspondence |
| C4            | ● NMP-analogs, cGMP, glucuronide-X, GS-conjugates | ● See C10                           | ● See C10                    | Partial correspondence |
| C5            | ● cGMP, organic anions, nucleotide analogs, GSH, GS-conjugates | ● See C1,2,3                        | ● See C1,2,3                 | Partial correspondence |
| C6            | ● ?                             | ● BTUD (E. coli)                     | ● Vitamin B12                | Prediction              |
| C7            | ● Chloride channel, GSH, GSSG, organic anions, bicarbonate | ● See C10                           | ● See C10                    | Multiple correspondences, 1/2 partially correct (i.e., GSG/GS-conjugates, cf. C10) |
| C8,9          | ● Sulfonylurea receptor         | ● MRP1,2 (A. thaliana)               | ● GS-conjugates, chlorophyll catabolites, glucuronides | Multiple correspondences, 1/3 correct (i.e., sulfonylurea receptor) |
|               |                                 | ● sur (D. melanogaster)              | ● Sulfonylurea receptor       |                          |
|               |                                 | ● See C1,2,3                        | ● See C1,2,3                 |                          |
| C10           | ● ?                             | ● YCF1 (S. cerevisiae)               | ● GS-conjugates, glutathionated Gd²⁺, bilirubin | Prediction |
| C11,12        | ● ?                             | ● See C1,2,3                        | ● See C1,2,3                 | Prediction |
| Note: BQ123 (cyclo(-D-Trp-D-Arp-Pro-D-Val-Leu-)Na⁺) is transported by the rat homolog of ABCG6.53 Like vitamin B12, it is cyclic but smaller. GSG/GSH, glutathione; GSSG, oxidized glutathione; NMP, nucleotide monophosphates; cGMP, cyclic guanine monophosphate. |
and included them in the profile as separate sequences. To detect them, we used the tool Repro.45 The resulting data set was then aligned by using DIALIGN,46 yielding a profile to be used by PSI-BLAST. We did not attempt to assemble the functional systems composed of several partners (e.g., dimerization partners); this task has been tackled on a large scale in case of bacterial systems,13 but extending their approach to eukaryotes would just add dimerization information that is not expected to provide useful pointers to functionality.

### TABLE II. (Continued)

<table>
<thead>
<tr>
<th>Human proteins a,b</th>
<th>Function/substances transported /human protein/</th>
<th>Nonvertebrate proteins of known function c,d</th>
<th>Function/substances transported /nonvertebrate protein/</th>
<th>Correspondence/prediction d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subfamily C (second halves)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1,2,3</td>
<td>● See first half</td>
<td>● mpr-1 (C. elegans)</td>
<td>● Cadmium ions, arsenite</td>
<td>Partial correspondence</td>
</tr>
<tr>
<td>C4</td>
<td>● See first half</td>
<td>● See C8,9</td>
<td>● See C8,9</td>
<td>No correspondence (see note)</td>
</tr>
<tr>
<td>C5</td>
<td>● See first half</td>
<td>● See C1,2,3</td>
<td>● See C1,2,3</td>
<td>Partial correspondence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● YCF1 (S. cerevisiae)</td>
<td>● See C1,2,3 first halves, BTP1 (S. cerevisiae)</td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>● ?</td>
<td>● See C8,9</td>
<td>● See C8,9</td>
<td>Prediction</td>
</tr>
<tr>
<td>C7</td>
<td>● See first half</td>
<td>● urtR (Synechocystis)</td>
<td>● LIVP (E. coli)</td>
<td>No correspondence</td>
</tr>
<tr>
<td>C8,9</td>
<td>● See first half</td>
<td>● sup (D. melanogaster)</td>
<td>● Sulfonylurea receptor</td>
<td>Correspondence</td>
</tr>
<tr>
<td>C10</td>
<td>● ?</td>
<td>● MRP1,2 (A. thaliana)</td>
<td>● See MRP1,2 (subfamily C first halves, column 4, row C8,9)</td>
<td>Prediction</td>
</tr>
<tr>
<td>C11,12</td>
<td>● ?</td>
<td>● See C5</td>
<td>● See C5</td>
<td>Prediction</td>
</tr>
</tbody>
</table>

Note: Like CFTR (C7), the sulfonylurea receptor (C8/9) may have transport competence. In that case, its transport functionality may still correspond to the one of C4. Interestingly, MRP5 of A. thaliana transports glucoronides,54 and it binds sulfonylurea.55 Sulfonylurea binding is, in fact, a feature of many ABC-related transporters.44

**Subfamily D**

D1,2,3

● (Very) long-chain fatty acids and/or their acyl-coenzyme A derivatives

● PXA2 (S. cerevisiae)

● Correlation/prediction

D4

● ?

● See D1,3 PXA1 (S. cerevisiae)

● Prediction

**Subfamily E**

There are no characterized nonvertebrate proteins in the whole subtree of this subfamily. Therefore, no correspondences or predictions can be made for ABCE1.

**Subfamily F (first/second halves)**

F1

● Interaction with eIF2

● GCN20 (S. cerevisiae)

● Activation of a kinase which phosphorylates eIF2α

● Correspondence

F2

● ?

● See F1

● See F1

● Prediction

F3

● ?

● See F1

● See F1

● Prediction

**Subfamily G**

G1,4

● Cholesterol, phosphatidylcholine

● See G2

● PDRs (S. cerevisiae)

● Scarlet, white (D. melanogaster)

● See G2

● Sterols, steroids, xenobiotics

● Eye pigment precursors (e.g., kynurenine?)

● Multiple correspondences; 2/3 correct

G2

● Sterols, steroids, steroid-conjugates, phosphatidylyserine, xenobiotics

● E21 (D. melanogaster)

● 20-hydroxyecdysone (steroid hormone)

● Partial correspondence

G5,8

● Lipids (sterols)

● See G1,4

● See G1,4

● Multiple correspondences; 2/3 correct

---

a Human/nonvertebrate correspondence was determined using the function transfer rule. The function transfer rule defines a subtree that includes the human protein(s) under consideration on the one hand (as listed in column 1) and the corresponding nonvertebrate protein(s) with known function on the other hand (as listed in column 3).

b Human proteins are denoted by the human classification system. Lists such as “C1,2,3” translate into C1, C2, C3, that is, ABCC1, ABCC2 and ABCC3. They are sorted numerically except that they start with the proteins for which experimental data are known. These proteins are underlined, and the functionality of the nonvertebrate proteins is compared with these, while it is predicted implicitly for the others that are not underlined.

c Certain entries refer to entries in other rows. For example, in column 3, an entry such as “see B7” in the row considering B6 means that the nonvertebrate proteins found for the former protein (B7 in this case) are the same as the ones found for the protein currently under consideration (B6 in this case; see Fig. 1).

d Correspondence/prediction terms follow Table III.

---

and included them in the profile as separate sequences. To detect them, we used the tool Repro.45 The resulting data set was then aligned by using DIALIGN,46 yielding a profile to be used by PSI-BLAST. We did not attempt to assemble the functional systems composed of several partners (e.g., dimerization partners); this task has been tackled on a large scale in case of bacterial systems,13 but extending their approach to eukaryotes would just add dimerization information that is not expected to provide useful pointers to functionality.
Retrieval-induced Phylogeny Environment

To estimate a phylogenetic tree of a protein (sub)family, the following steps are performed using the original RiPE algorithm, given a profile of protein sequences as query:

1. Use the aligned (sub)family sequences as a query profile (jumpstart alignment) for a PSI-BLAST database search.
2. Convert the PSI-BLAST report into an alignment, stacking the high-scoring segments of the hits to the query profile via MVIEW.
3. Filter sequences that do not contain ABC protein family motifs.
4. Filter sequences that are almost identical (95% or more) to another sequence, or that are part of another sequence, unless the organism/species is different.
5. Rename sequences by converting the database definition lines into a standardized human-readable format.
6. Determine a probability of homology (more accurately, an E-value) threshold in order to retain only sequences that belong to the (sub)family in question.
7. Realign the sequences.
8. Perform a phylogenetic analysis.

Step 7 can be computationally very expensive, even if recently published fast alignment methods are used (e.g., MAFFT). Therefore, we omitted step 7 from the analyses in this article. We found that the alignment obtained by stacking is even more reliable without the realignment step, speeding up the analysis and ignoring only data of doubtful homology. Moreover, in contrast to the original RiPE pipeline, we refrained from modifying the PSI-BLAST parameters to extend the high-scoring segments that make up the hits. Finally, we improved automation of the RiPE pipeline by making expert knowledge of conserved motifs redundant. In the original version, fragmental sequences were filtered out by requiring the presence of homologous residues at pre-specified sites of high conservation. We now use PRATT to identify these sites automatically for step 3. Step 8 is now performed using the QuickTree implementation of the neighbor-joining algorithm. Note that the second copies of ABC domains in database sequences are retrieved as separate hits, because all sequences in the query profile feature only one ABC (α or β). The running duration of RiPE is dominated by the tree reconstruction algorithm using QuickTree, which currently takes about
TABLE III. Correspondences and Predictions, Definition of Termsa

<table>
<thead>
<tr>
<th>Case</th>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Correspondence</td>
<td>Close match between functionalities of the human and the nonvertebrate proteins.</td>
</tr>
<tr>
<td>2</td>
<td>Partial correspondence</td>
<td>The functionality of the nonvertebrate protein(s) is a subset of functionality of the human protein(s), and/or vice versa.</td>
</tr>
<tr>
<td>3</td>
<td>Multiple correspondences, x/y correct</td>
<td>There are y nonvertebrate proteins with known functionality in the sister group, and for x of these, functionality corresponds with the human protein(s).</td>
</tr>
<tr>
<td>4</td>
<td>Prediction</td>
<td>The functionality of the human protein(s) is unknown and predicted from the nonvertebrate protein(s).</td>
</tr>
</tbody>
</table>

a Cases 2 and 3 are also called “weak correspondences.” Boldface, underlining, etc. are used as a convention for easier recognition in Table II.

99% of the time. The modified RiPE pipeline is available from the authors on request.

Function Predictions and Correspondences

Functional annotations relevant for the investigation of the human ABC proteins were compiled for this article from Holland et al.,27 Dean,51 the Human ABC Transporter website (http://nutrigene.4t.com/humanabc.htm), PubMed searches,44 and the genomic databases listed in Table SVI in the Supplementary Material. The annotations are listed in the second and fourth column of Table II. They are all backed up by experimental data; we did not include any predicted functionality (based on expert assessment or computation). We did not consider mouse protein annotations, because these are few in number and very closely related to human ABC protein inventories.31 At the time of writing, there were no experimentally validated Fugu annotations for ABC transporters. Consequently, only annotation data regarding nonvertebrate sequences were used. Functional annotations were considered whenever they could be used to formulate predictions/correspondences for the human sequences, according to the following rule for transfer of functionality (the function transfer rule), given the phylogenetic tree of the sequences of a (sub)family:

For each human protein with a (possibly unknown) functionality f, we descend the tree node by node toward the root. Each such node corresponds to the common ancestor of the subtree from which we come, and of the other subtree (consisting of the sister group).

If the sister group includes a nonvertebrate protein (or proteins) with experimentally verified functionality g, h, and so on, we terminate and make a prediction (or check functional correspondence of f versus g, f versus h, and so on, if functionality f is known).

Otherwise, we take note of any human proteins in the sister group and generalize functionality f as necessary, if the functionalities of these human proteins do not match f. We continue descending the tree. Note that any prediction still made also applies to any noted human protein.

An example of the function transfer rule can be found in Figure 1. Consider the human ABCB7 protein. No annotations for ABCB7 were found. Consequently, for ABCB6, we follow the light-gray path (orange online), and in the sister group we find the proteins just mentioned, including ATM3 and ATM1. Again, this corresponds to what is known about ABCB6 (see Table II). Some further discussion of the function transfer rule can be found in the Supplementary Material.

Tree Simplification

The species tree of Figure S2 in the Supplementary Material is based on the NCBI taxonomy.44 Its structure is widely corroborated except that some researchers believe that worm and fly clades belong together (eucaryotes hypothesis, see Blair et al.57 for a recent discussion), and that the position of Encephalitozoon cuniculi is not well-established.58 Given the species tree and the classification scheme of Table SV in the Supplementary Material, the tree obtained by RiPE (see the Supplementary Material) was simplified manually by the following set of rules yielding Figure 2:

1. Monophylum compression. Subtrees with sequences classified alike that follow the species phylogeny, or that belong to a single species, were replaced by a single label designating the (group of) species to which they belong (e.g., “Eukarya F1,” “Bilateria F1,” “Bact. (FAE).” The label “Worm/Fly” is used instead of “Bilateria” if only worm and fly (but not H. sapiens) are included as species designations. If all sequences are unclassified (e.g., Caenorhabditis elegans sequences without common gene names), they are summarized analogously (e.g., “Worm”).

2. Unclassified sequences mixed with classified ones are assumed to have the same label as the classified ones, and they are ignored except that they contribute their species designation to the labels.

3. Mouse orthologs to human proteins are not mentioned, since they cluster with their human counterparts, except for the single case of mouse Abcc2, which clusters with plant sequences. Moreover, single Fugu labels are suppressed.

4. We distinguish between Anopheles gambiae and Droso- phila melanogaster only as terminal leaves. We do not distinguish if compression can be achieved. Moreover, if sequences from both species occur in the same subtree, the annotation is “Fly.” The distinction is made because
A. gambiae has a distinct inventory of ABC proteins which are often shared with bacteria, but not with D. melanogaster.

5. We do not distinguish between Arabidopsis thaliana and Oryza sativa proteins. Both species are designated as “Plant.” The reason is that O. sativa ABC proteins have not yet been experimentally characterized.

6. We do not distinguish between Saccharomyces cerevisiae and Schizosaccharomyces pombe. Both species are
designated as “Yeast.” Both species have a similar ABC protein inventory.

7. *Paraphylum compression*. As long as consecutive edges in a subtree make up a backbone to which subsubtrees are attached that feature sequences classified alike and that belong to the same species/clade, these edges are deleted, and a single subtree that subsumes all subsubtrees is introduced. For example, this rule is used to write a single label “Fly/Worm” replacing the succession of subtrees marked in gray in Figure S3 in the Supplementary Material.
RESULTS AND DISCUSSION

Using RiPE to analyze all ABC proteins of 20 model organisms (see Materials and Methods section), we built a tree of 1138 sequences, the essence of which is presented in Figure 2, including an analysis of functional invariance. The raw version of the tree was rendered by njplot and as described in the Supplementary Material, the tree is almost complete, and only 11 of 264 members known to belong to a specific ABC subfamily are included in the subtree of a different subfamily.

Overall Tree Structure

The tree as summarized in Figure 2 features 11 subtrees, eight of which correspond to the seven known human ABC subfamilies designated by letters A to G (cf. Table I; the C subfamily is divided into two subtrees corresponding to the first and second halves). Corroborating a birth-and-death history of many ABC genes, the model organism tree shown in Figure S2 in the Supplementary Material is rarely reflected in the protein (gene) subtrees. However, there are exceptions, for example, in case of the F subfamily, where the gene tree follows the species tree at least from the “Bilateria” level onwards. The trees for subfamilies A, E, and F are composed of two subtrees each, one for each half of the tree (full-transporter, see the Introduction) or α domain arrangement. Dean et al. present a tree of the human sequences only that features the same structure in case of the F subfamily. Their E subfamily, however, is incorrectly represented by a single half. Their subfamily A tree groups the first halves of A1 to A4, A7, and A12 with all other second halves, with high resampling (bootstrap) support. Our tree is more plausible because it features one subtree for each half.

For most subfamilies, bacterial sequences are placed close to the subtree root. Moreover, the domain arrangement of the eukaryote sequences is reflected by the domain arrangement of their bacterial neighbors in the tree (see Fig. 2 and Table I). Thus, we assume that before the kingdoms split apart, ABC proteins diversified into subfamilies, and many of the genes contained both the ABC cassette (“a”) and the transmembrane region (“t”). As in Dean et al., the Drosophila ABCH sequences with domain arrangement “at” are located at the root of an ABCA subtree featuring proteins with domain arrangement “at”; the A subfamily features no domain arrangement that is fixed across kingdoms. This leads to the hypothesis that fusion of the A subfamily protein domains took place rather late in evolutionary time. The only subtree with no bacterial members featuring only eukarya and archaea corresponds to the E subfamily. The human ABCE protein is implicated in RNase L inhibition. This observation is in agreement with the hypothesis that information-processing proteins in eukaryotes are of archaeal origin.

The three subtrees labeled “Various” feature a few eukaryotic proteins of Anopheles gambiae, Arabidopsis thaliana (subfamily NAP as defined by Sanchez-Fernandez et al. to designate all proteins that could not be classified into a known group), Plasmodium falciparum, and Encephalitozoon cuniculi. Often, these proteins cluster with bacterial sequences at a resampling (bootstrap) support level of more than 90%. They may be lost in the other eukaryotic lineages or may have been transferred horizontally. The latter idea, in the form of the endosymbiont hypothesis (see Materials and Methods section), is corroborated by the observation that in many cases, proteins of Synechocystis sp. cluster with A. thaliana, while those of R. prowazekii cluster with A. gambiae.

Function Predictions and Correspondences

We used the phylogenetic tree just described and the function transfer rule (see Materials and Methods section) to obtain functional predictions for the human sequences, based only on the knowledge of nonvertebrate proteins. The function transfer rule defines a subtree that features a functionality that is supposedly shared by all proteins in that subtree. The rule relies on the parsimony assumption that a certain functionality is maintained within a gene (sub)family despite speciation and duplication events, and that phenomena such as the recruitment of novel functionality (cf. Benner et al.) are rare events. If, at the time of writing, something was known about the functionality of the human sequences included in the subtree, we checked the correspondence of the prediction with the known human annotation. Table II provides a detailed list of predictions and correspondences (cf. the definitions in Table III) sorted by subfamily.

In the case of subfamily A (with respect to the first halves only) and subfamily E, no nonvertebrate homologs with known functionality were found at the time of writing, and functionality transfer is impossible. However, the situation is not as bleak as described by Dean et al., who negate the utility of comparative analysis in most cases based on D. melanogaster and C. elegans data only. For the other subfamilies, we observed five correspondences, and 14 weak correspondences. We obtained 10 predictions; in only three cases there was no correspondence between the functional annotation and the known function of the protein. We expect that in most of these cases, the error lies in our transfer rule, but an experimental test of the predicted functionality of the human protein may be worthwhile nevertheless.

Among the correspondences, we found that functionality of the A subfamily predicted from the nonvertebrate protein can be subsumed by “transport of complex lipids,” while the human proteins transport a broader range of substances. Within the B subfamily, the nonvertebrate sequences give valid hints at transport of iron–sulfur cluster protein precursors and peptides by the human members. For the C subfamily, sulfonyleurea receptor activity and transport of glutathione conjugates are predicted successfully. The chloride channel ABCC7 clusters, among others, with a glutathione conjugate transporter (yeast YCF1, which is closely related to ABCC10 in the

†Mouse and Fugu sequences are included in the tree, but mouse is too close to human to render interesting predictions, and Fugu features no annotations with experimental data yet.
tree), and it has been shown that ABCC7 can act as a transporter of glutathione (cf. Table II). The remaining ABC transporter subfamilies, ABCD and ABCG, feature correspondences with respect to the transport of (very) long-chain fatty acids and steroids/sterols. Subfamily F members interact with certain types of transcription factors, in the case of both vertebrate and nonvertebrate proteins. Further research will clarify which of the predictions listed in Table II turn out to be valid; in the future we plan to pursue this analysis with a larger array of model organisms, or, possibly, using the entire “nonredundant” database at GenBank.

**Comparison with Function Prediction by Direct Inspection of BLAST Reports**

We compared the application of the function transfer rule to the RiPE-based phylogenetic tree with the direct inspection of BLAST search results. We investigated all eight cases of “correspondence” or “no correspondence,” as well as all six cases of “multiple correspondences, x/y correct” (see Tables II and III). We conducted a BLAST search with every functionally characterized human protein to be compared, using standard parameters and an E-value of 10^-50. In the resulting lists of hits, we noted the first three functionally characterized nonhypothetical vertebrate proteins. This characterization matched ours completely in seven cases. In five cases, our characterization as given in Table II is superior as follows. The BLAST search did not produce any predictions for ABCF1, and it produced incorrect predictions for the sulfonylurea receptor ABCC8/9 (first half, as well as the second half; hits are yeast BPT1 (for the first half) and yeast YCF1, *Arabidopsis* MRP1/2, yeast YBT1 (for the second half)), and for ABCG1 and ABCG5/8 (in both cases, the only characterized hit was the *Drosophila* white protein, which does not hint at the function of human ABCG1/5/8). The BLAST search was superior in two cases: ABCC4 and ABCC7 (second halves) were characterized more precisely as transporting GS-conjugates, by way of yeast YCF1, which was the only hit in both cases.

**Focus on Eukaryotic Protein Data**

The aim of this study was to demonstrate the feasibility of predicting the function of human ABC proteins. We used in our analyses well-annotated organisms and all completely sequenced eukaryotes available at the time of writing, because valid function transfer from eukaryotes to human is more likely to be successful than transfer from bacteria or archaea. However, it would also be feasible to conduct the first part of our analysis (homology search and tree reconstruction) with many more organisms, in particular more bacteria and archaea. The second part (cartoon tree generation and function transfer) would be more difficult to expand because these steps are performed manually. Although it is possible to automate simplification of phylogenetic trees (Lott et al., in preparation), collecting functional annotations from a large set of databases and publications, and interpreting their relationship, is at present a mostly manual task, because function is not a concept that can be automated adequately at present. Overall, for the task of function prediction of human proteins, the selection of taxa used in this study was sufficient to demonstrate feasibility of the pipeline.

Our tree closely matches the essence of previously published trees, which were often based on more taxa, even though the authors of these studies did not analyze eukaryotic, bacterial, and archaeal sequence data simultaneously on a large scale, as we did. Thus, we are confident that adding more data would not have changed the topology of the tree and the conclusions derived from it, in particular with respect to function. However, we expect that inclusion of more organisms will allow more, and more precise, predictions, in particular if more organisms are annotated experimentally in the future. This may, for example, allow function prediction for the two ABC subfamilies for which we could not obtain nonhuman annotations.

**CONCLUSIONS**

We have shown that RiPE combined with the function transfer rule can be used to provide good functional annotation of human ABC proteins. We are confident that this annotation can be improved if data from even more species are included. Our focus on human proteins is determined by the profile of human sequences with which we started the database search. However, we could also have focused on other organisms simply by using another set of sequences as the profile. Therefore, RiPE is also a suitable tool for obtaining functional annotation for ABC proteins in the newly assembled protein data sets of other organisms. Future work will include further automation and formalization of the tree simplification and visualization steps, and of the comparison of functional annotations using a detailed ontology.6 The multidomain problem of protein functionality estimation and comparison is reduced by RiPE to the identification and analysis of homologous partial sequences without domain repeats. In the present work, we ignore the possibility that different (sub)domains (e.g., ATP-binding cassette vs transmembrane region) may give rise to different functional annotations. We are not aware of any such problems in the case of ABC proteins. Generalization of our RiPE pipeline to the automated analysis of arbitrary protein families would nevertheless benefit from a solution to this problem.

**ACKNOWLEDGMENTS**

P. Cullen and S. Lorkowski were participants of the project “Macrophage Function and Stability of the Atherosclerotic Plaque” (QLG2-CT-1999-01007) supported by the European Union.65 We thank Holger Wagner, Christian Rückert, and Elie Dassa for critical reading and valuable comments on the manuscript.

**REFERENCES**

2. Benner SA, Chamberlin SG, Liberles DA, Govindarajan S, Knecht L. Functional inferences from reconstructed evolutionary biology involving rectified databases—an evolutionarily grounded ap-