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

Georg Fuellen,[1,2*] Michael Spitzer,[1,2,3] Paul Cullen,[4] and Stefan Lorkowski[4,5]
[1]*Department of Medicine, AG Bioinformatics, University of Münster, Münster, Germany*
[2]*Division of Bioinformatics, Biology Department, University of Münster, Münster, Germany*
[3]*Integrated Functional Genomics, Interdisciplinary Center for Clinical Research, Münster, Germany*
[4]*Leibniz Institute of Arteriosclerosis Research, University of Münster, Münster, Germany*
[5]*Institute 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.**

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 hits[3,4]; 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 rapidity with which an alignment can be generated directly from the results of a profile database search, and the speed of newer phylogeny software[22] 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 database[26] 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.[28] 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

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

*Correspondence to: Georg Fuellen, Department of Medicine, c/o Division of Bioinformatics, Schlossplatz 4,1. Etage, 48149 Münster, Germany. E-mail: fuellen@uni-muenster.de

Received 15 November 2004; Accepted 24 March 2005

Published online 27 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20616

TABLE I. Domain Arrangements of Proteins Found (Listed by Subtree; cf. Introduction and Figure 2)

Subfamily[a]	Functional class[b]	Domain arrangements	
		Eukarya	Bacteria, Archaea
A; H	DRI/DRA	ta, taτα; at	a, aα, aαt, ta
B	DPL	ta, taτα	ta
C (first half)	OAD (DPL, CCM, ISVH, NO)	ta	ta, a
C (second half)	OAD (ABCY)	ταta	ta, a
D	FAE	ta	ta
E	RLI	aα	aα
F	ART	aα	aα
G	EPD	at, atατ	at

[a] Klein, Sarkadi, and Varadi[30]; Human ABC Transporter website (http://nutrigene.4t.com/humanabc.htm); Dean, Rzhetsky, and Allikmets.[31]
[b]Dassa and Bouige[32]; classes of bacterial sequences found in a subfamily that do not match are enclosed in parentheses.

C-terminal region that includes residues typical of the ABC subfamily G (see Table I for a list of ABC subfamily designations).[29] 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.[33] 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.[34] 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 α) and two less similar transmembrane regions (symbols t and τ), mostly in the order "taτα." Thus, "taτα" is a short form of the standard notation $TMD_1ABC_1TMD_2ABC_2$, 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.[33] 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,[32] which is not restricted to bacterial ABC proteins. As noted by Tomii and Kanehisa,[35] 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,[36] and the one by Tam and Saier.[37] 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.[35] 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 SII 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*,[38,39] presumed close relatives of the endosymbionts that are believed to be the ancestors of mitochondria (*Rickettsia prowazekii*[40–42]) and chloroplasts (*Synechocystis* sp.[43]) 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[44] 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 "τα" or "ατ") prevents achievement of consistent overall alignment since sequences with just one copy may align to either one of the two copies in the aα, taτα, or atατ arrangement. For this reason, we deleted the second copies

TABLE II. Human and Nonvertebrate ABC Proteins, Function Predictions, and Correspondence of Annotations (references and additional data are given in Table SI of the Supplementary Material)

Human proteins[a,b]	Function/substances transported (human protein)[c]	Nonvertebrate proteins of known function[a,c]	Function/substances transported (nonvertebrate protein)[c]	Correspondence/prediction[d]
Subfamily A (first halves)				
There are no characterized nonvertebrate proteins in the whole subtree of the first halves of the proteins of this subfamily. Therefore, no correspondences or predictions can be made.				
Subfamily A (second halves)				
A1,2,3,4,7,12	• Phospholipids • Estramustine • Lung surfactant preprocessing • *N*-retinylidene-phosphatidyl-ethanolamine	• drrA (*M. tuberculosis*)	• Phthiocerol dimycocerosate (complex lipid), doxorubicin	*Partial correspondence*
A8,5,6,9,10	• Xenobiotics [e.g., estradiol-β-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.*, phospatidylcholine, colchicines	• mdr49 (*D. melanogaster*) • mdr65 (*D. melanogaster*) • pgp-1 (*C. elegans*)	• Colchicines • α-amanitin • Rhodamine 123	Multiple correspondences, 1/3 correct (i.e., colchicines)
B11	• Bile salts, paclitaxol	• See B1,4,5	• See B1,4,5	No correspondence
B2,3,9	• Peptides	• See B8,10 • PMD1 (*S. pombe*) • MDL1 (*S. cerevisiae*)	• See B8,10 • Leptomycin B • Peptides	Multiple correspondences, 1/3 correct (i.e., peptides)
B6	• Iron metabolism	• See B7	• See B7	*Partial correspondence* (see note)
B7	• Fe/S cluster metabolism	• ATM3 (*A. thaliana*) ATM1 (*S. cerevisiae*)	• Fe/S proteins	**Correspondence**
B8,10	• ?	• MSBA (*E. coli*) • MDR1 (*A. thaliana*)	• Lipid A, phospholipids, lipopolysaccharides • Auxin (tryptophan derivative)	Prediction
Note: Involvement in iron metabolism is considered to correspond weakly to involvement in Fe/S cluster metabolism.				
Subfamily C (first halves)				
C1,2,3	• Sulfate, glutathione and glucoronide conjugates of organic arsenite	• MRP3,4,5 (*A. thaliana*)	• GS-conjugates, chlorophyll catabolites, glucuronides	**Correspondence**
C4	• NMP-analogs, cGMP, glucoronide-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 B_{12}	Prediction
C7	• Chloride channel, GSH, GSSG, organic anions, bicarbonate	• See C10 • YOR1 (*S. cerevisiae*)	• See C10 • Mycins, phospholipids	Multiple correspondences, 1/2 partially correct (i.e., GS/GS-conjugates, cf. C10)
C8,9	• Sulfonylurea receptor	• MRP1,2 (*A. thaliana*) • sur (*D. melanogaster*) • See C1,2,3	• GS-conjugates, chlorophyll catabolites, glucuronides • Sulfonylurea receptor • See C1,2,3	Multiple correspondences, 1/3 correct (i.e., sulfonylurea receptor)
C10	• ?	• YCF1 (*S. cerevisiae*) BPT1 (*S. cerevisiae*)	• GS-conjugates, gluthathionated Cd^{2+}, bilirubin	Prediction
C11,12	• ?	• See C1,2,3	• See C1,2,3	Prediction
Note: BQ123 (cyclo(-D-Trp-D-Asp-Pro-D-Val-Leu-)Na^{+}) is transported by the rat homolog of ABCC6.[53] Like vitamin B_{12}, it is cyclic but smaller. GS/GSH, glutathione; GSSG, oxidized glutathione; NMP, nucleotide monophosphates; cGMP, cyclic guanine monophophate.				

TABLE II. (Continued)

Human proteins[a,b]	Function/substances transported (human protein)[c]	Nonvertebrate proteins of known function[a,c]	Function/substances transported (nonvertebrate protein)[c]	Correspondence/prediction[d]
Subfamily C (second halves)				
C1,2,3	• See first half	• mrp-1 (*C. elegans*)	• Cadmium ions, arsenite	*Partial correspondence*
C4	• See first half	• See C8,9	• See C8,9	No correspondence (see note)
C5	• See first half	• See C1,2,3 • YCF1 (*S. cerevisiae*) BPT1 (*S. cerevisiae*)	• See C1,2,3 • See YCF1 and BTP1 above (subfamily C first halves, column 4, row C10)	*Partial correspondence*
C6	• ?	• See C8,9	• See C8,9	Prediction
C7	• See first half	• urtE (*Synechocystis*) • BRAG (*Synechocystis*) • LIVF (*E. coli*)	• Urea • Neutral amino acids • Branched-chain amino acids	No correspondence
C8,9	• See first half	• sur (*D. melanogaster*)	• Sulfonylurea receptor	**Correspondence**
C10	• ?	• MRP1,2 (*A. thaliana*)	• See MRP1,2 above (subfamily C first halves, column 4, row C8,9)	Prediction
C11,12	• ?	• See C5	• See C5	Prediction
Note: Like CFTR (C7), the sulfonylurea receptor (C8/C9) 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] Sulfonnylurea binding is, in fact, a feature of many ABCC-related transporters.[56]				
Subfamily D				
D1,3,2	• (Very) long-chain fatty acids and/or their acyl-coenzyme A derivatives	• PXA2 (*S. cerevisiae*)	• (Very) long-chain fatty acids and/or their acyl-CoA derivatives	**Correspondence**
D4	• ?	• See D1,3 PXA1 (*S. cerevisiae*)	• See D1,3	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, phosphatitylcholine	• 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, phosphatidylserine, xenobiotics	• E23 (*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.

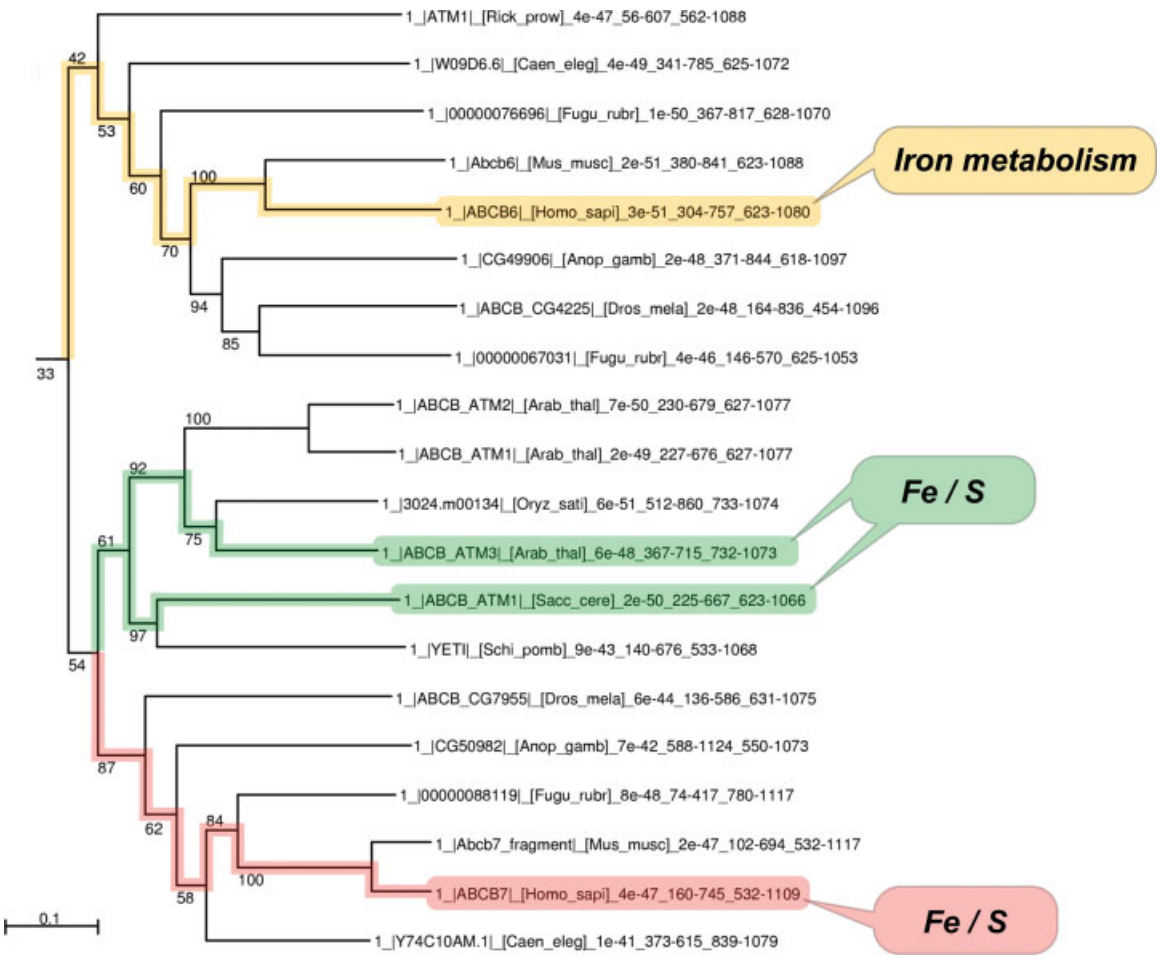

Fig. 1. Application of the function transfer rule in a subtree of subfamily B (see text for details). (Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.)

Retrieval-induced Phylogeny Environment

To estimate a phylogenetic tree of a protein (sub)family, the following steps are performed using the original RiPE algorithm,[33] 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.[47]
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[48]). 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.[33] 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 prespecified sites of high conservation. We now use PRATT[49] to identify these sites automatically for step 3. Step 8 is now performed using the QuickTree implementation of the neighbor-joining algorithm.[22,50] 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 (a or α). 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 Terms[a]

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

[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 based on experiments are known for the related mouse, *Fugu*, worm, or fly proteins. Descending the subtree along the dark-gray path (red online), we finally find a sister tree that features *Arabidopsis* ATM3 and *Saccharomyces* ATM1 proteins known to be involved in iron–sulfur (Fe/S) cluster protein metabolism. Function transfer is successful in this case: Human B7 is known to possess this functionality (see Table II). 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 (ecdysozoa 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 no *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 Fugue labels are suppressed.
4. We distinguish between *Anopheles gambiae* and *Drosophila 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

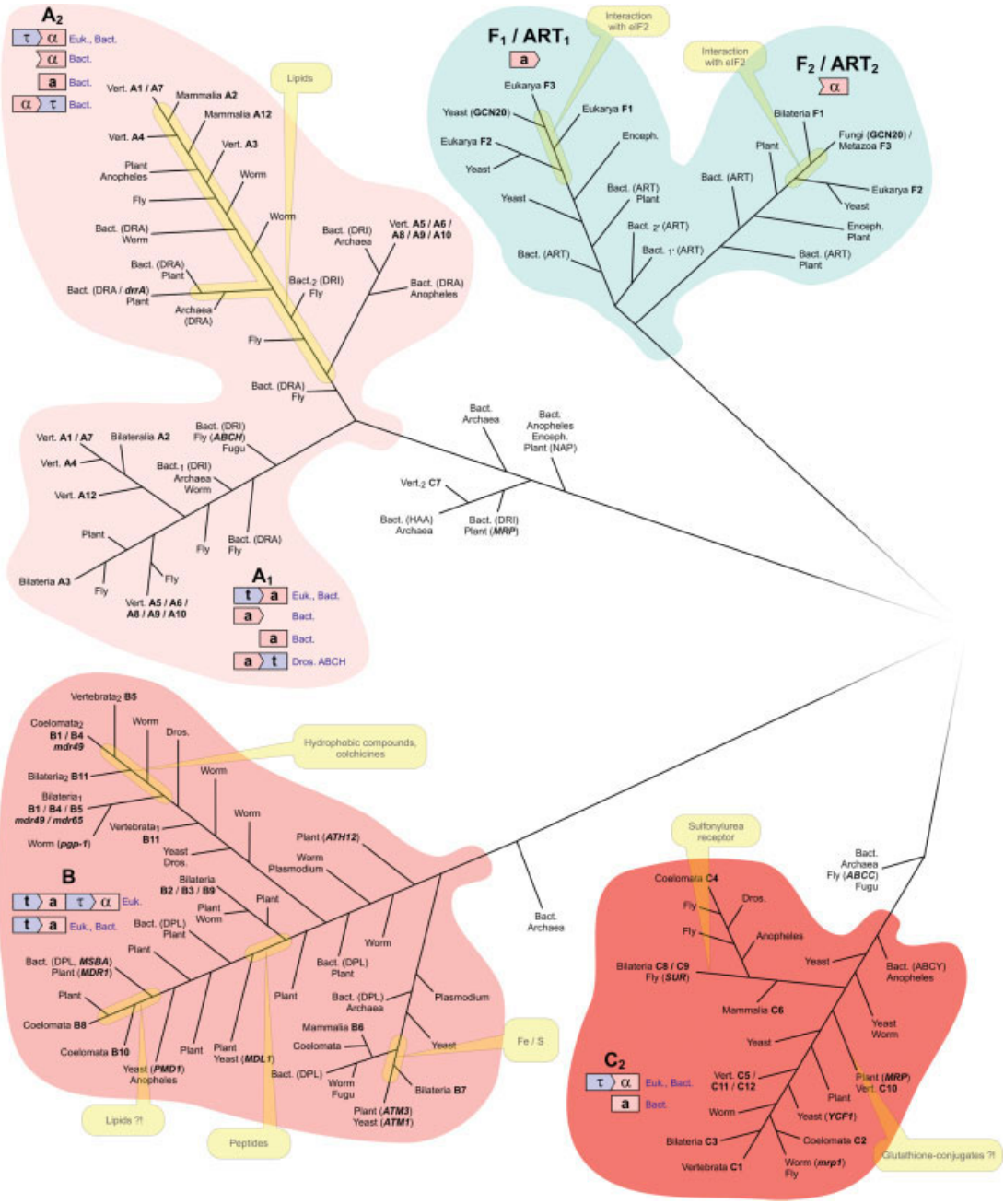

Fig. 2. Overall structure of the RiPE tree of the ABC protein family. The subfamily trees are delineated by color. They are labeled by family designation letters (A, B, …, G) and by the code of Dassa and Bouige[32] (cf. Table I) where customary. Subscripts 1 and 2 (as in F_1) denote the first or second domain. Subscripts 1′ and 2′ (as in Bact.$_1$·(ART) in the F_1 subtree) denote a first or second domain resulting from another, subsequent duplication. The domain arrangements ("ta," "aα," etc., as described in the text) are placed next to the subfamily label. The human ABC proteins (**A1**, **A2**, …, **B1**, **B2**, etc.) are typed in boldface. Gene names of nonhuman proteins (e.g., ***PXA1***, ***PXA2***) are given in italics and boldface. Bacterial or archaeal proteins classified by Dassa and Bouige[32] are labeled by the corresponding code [e.g., "Bact. (ART)"], if a common clade exists. If they appear as single sequences, the gene name is preferred. Finally, a balloon is used to specify a subtree for which shared functionality is predicted by the function transfer rule. The prediction is based on the annotated nonvertebrate sequence(s) in the subtree. Functionality is also transferred to human proteins via a path that goes further toward the root of the tree, if there are no closer related nonvertebrate proteins with known functionality. Yellow ovals surround the edges along which functionality is supposed to be invariant in that case.

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

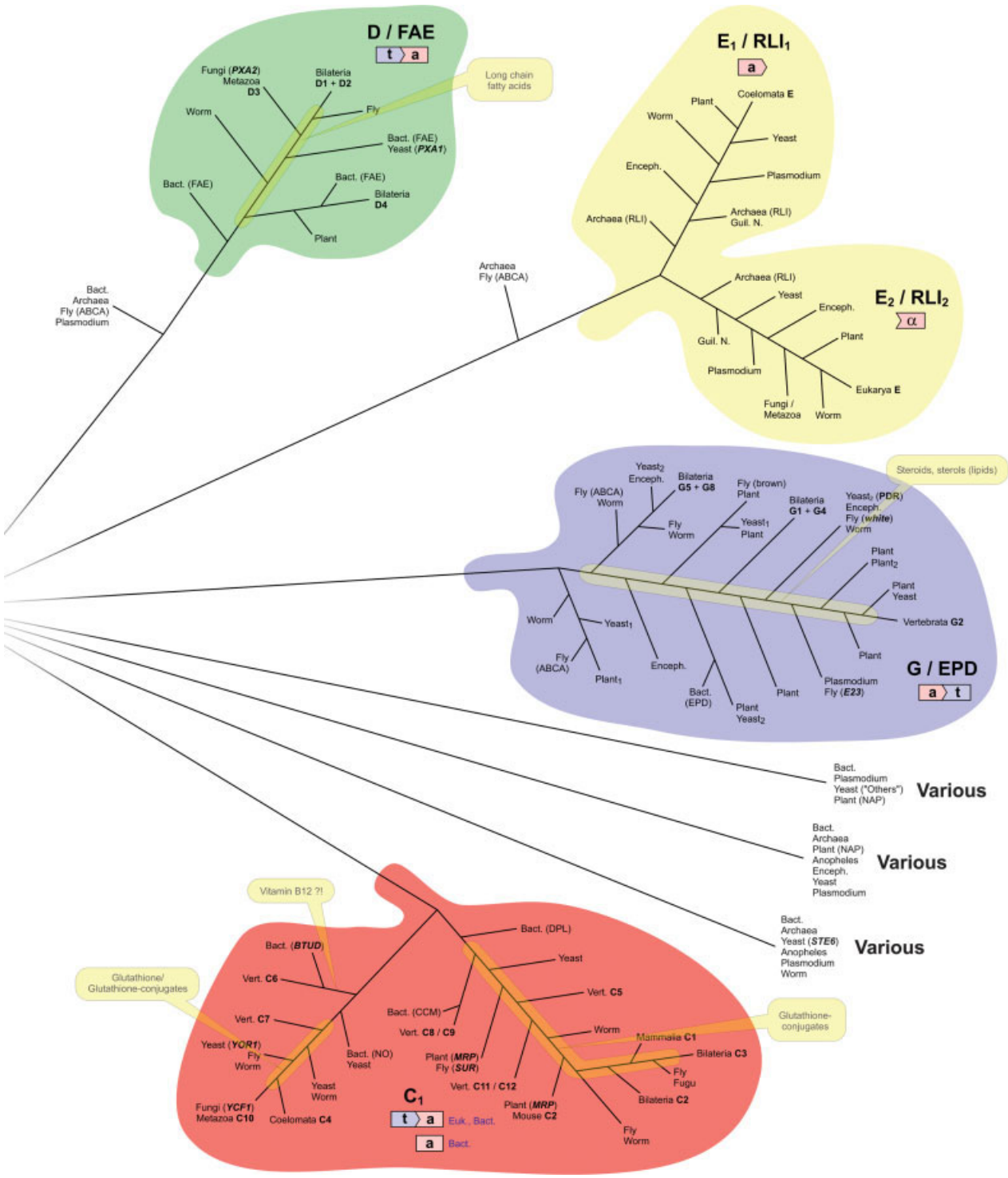

Figure 2. (Continued)

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[59] 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,[31] 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 taτα (full-transporter, see the Introduction) or aα domain arrangement. Dean et al.[31] 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.,[31] the *Drosophila* ABCH sequences with domain arrangement "at" are located at the root of an ABCA subtree featuring proteins with domain arrangement "ta"; 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.[60] This observation is in agreement with the hypothesis that information-processing proteins in eukaryotes are of archaeal origin.[61]

The three subtrees labeled "Various" feature a few eukaryotic proteins of *Anopheles gambiae*, *Arabidopsis thaliana* (subfamily NAP as defined by Sanchez-Fernandez et al.[62] 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.[2]) 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.,[31] 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, sulfonylurea 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.[63] The multidomain problem of protein functionality estimation and comparison[64] 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

1. Copley RR, Letunic I, Bork P. Genome and protein evolution in eukaryotes. Curr Opin Chem Biol 2002;6:39–45.
2. Benner SA, Chamberlin SG, Liberles DA, Govindarajan S, Knecht L. Functional inferences from reconstructed evolutionary biology involving rectified databases—an evolutionarily grounded ap-

proach to functional genomics. Res Microbiol 2000;151:97–106. See also references therein.
3. Da Silva AC, Ferro JA, Reinach FC, Farah CS, Furlan LR, Quaggio RB, Monteiro-Vitorello CB, Van Sluys MA, Almeida NF, Alves LM, do Amaral AM, Bertolini MC, Camargo LE, Camarotte G, Cannavan F, Cardozo J, Chambergo F, Ciapina LP, Cicarelli RM, Coutinho LL, Cursino-Santos JR, El-Dorry H, Faria JB, Ferreira AJ, Ferreira RC, Ferro MI, Formighieri EF, Franco MC, Greggio CC, Gruber A, Katsuyama AM, Kishi LT, Leite RP, Lemos EG, Lemos MV, Locali EC, Machado MA, Madeira AM, Martinez-Rossi NM, Martins EC, Meidanis J, Menck CF, Miyaki CY, Moon DH, Moreira LM, Novo MT, Okura VK, Oliveira MC, Oliveira VR, Pereira HA, Rossi A, Sena JA, Silva C, de Souza RF, Spinola LA, Takita MA, Tamura RE, Teixeira EC, Tezza RI, Trindade dos Santos M, Truffi D, Tsai SM, White FF, Setubal JC, Kitajima JP.Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. Nature 2002;417:459–463.
4. Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF Jr, Woo L, Chen Y, Paulsen IT, Eisen JA, Karp PD, Bovee D Sr, Chapman P, Clendenning J, Deatherage G, Gillet W, Grant C, Kutyavin T, Levy R, Li MJ, McClelland E, Palmieri A, Raymond C, Rouse G, Saenphimmachak C, Wu Z, Romero P, Gordon D, Zhang S, Yoo H, Tao Y, Biddle P, Jung M, Krespan W, Perry M, Gordon-Kamm B, Liao L, Kim S, Hendrick C, Zhao ZY, Dolan M, Chumley F, Tingey SV, Tomb JF, Gordon MP, Olson MV, Nester EW. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science 2001;294:2317–2323.
5. Remm M, Sonnhammer E. Classification of transmembrane protein families in the *Caenorhabditis elegans* genome and identification of human orthologs. Genome Res 2000;10:1679–1689.
6. Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. J Mol Evol 2001;52:540–542.
7. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 2001;314:1041–1052.
8. Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, Holt I, Liang F, Quackenbush J. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res 2002;12:493–502.
9. Zmasek CM, Eddy SR. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 2002;3:14.
10. Jensen LJ, Ussery DW, Brunak S. Functionality of system components: conservation of protein function in protein feature space. Genome Res 2003;13:2444–2449.
11. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003;4:41.
12. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 2001;17:847–848.
13. Quentin Y, Chabalier J, Fichant G. Strategies for the identification, the assembly and the classification of integrated biological systems in completely sequenced genomes. Comput Chem 2002;26: 447–457.
14. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, Cherry JM, Botstein D. Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science 1998;282:2022–2028.
15. Eisen JA, Hanawalt PC. A phylogenomic study of DNA repair genes, proteins, and processes. Mutat Res 1999;435:171–213.
16. Johnson JM, Church GM. Predicting ligand-binding function in families of bacterial receptors. Proc Natl Acad Sci USA 2000;97: 3965–3970.
17. Sicheritz-Ponten T, Andersson SG. A phylogenomic approach to microbial evolution. Nucleic Acids Res 2001;29:545–552.
18. Eisen JA, Wu M. Phylogenetic analysis and gene functional predictions: phylogenomics in action. Theor Popul Biol 2002;61: 481–487.
19. Joost P, Methner A. Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands. Genome Biol 2002;3:RESEARCH0063.
20. Leipe DD, Wolf YI, Koonin EV, Aravind L. Classification and evolution of P-loop GTPases and related ATPases. J Mol Biol 2002;317:41–72.
21. Liberles DA, Wayne ML. Tracking adaptive evolutionary events in genomic sequences. Genome Biol 2002;3:REVIEWS1018.
22. Howe K, Bateman A, Durbin R. QuickTree: building huge neighbour-joining trees of protein sequences. Bioinformatics 2002;18: 1546–1547.
23. Rannala B, Huelsenbeck JP, Yang Z, Nielsen R. Taxon sampling and the accuracy of large phylogenies. Syst Biol 1998;47:702–710.
24. Ohta T. Evolution of gene families. Gene 2000;259:45–52.
25. Ponting CP, Birney E. Identification of domains from protein sequences. Methods Mol Biol 2000;143:53–69.
26. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D. ProDom: automated clustering of homologous domains. Brief Bioinform 2002;3:246–251.
27. Holland IB, Cole S, Kuchler K, Higgins C. ABC-transporters from bacteria to man. Amsterdam: Academic Press; 2002.
28. Gracy J, Argos P. DOMO: a new database of aligned protein domains. Trends Biochem Sci 1998;23:495–497.
29. Lorkowski S, Cullen P. ABCG subfamily of human ATP-binding cassette proteins. Pure Appl Chem 2002;74:2057–2081.
30. Klein I, Sarkadi B, Varadi A. An inventory of the human ABC proteins. Biochim Biophys Acta 1999;1461:237–262.
31. Dean M, Rzhetsky A, Allikmets R. The human ATP-binding cassette (ABC) transporter superfamily. Genome Res 2001;11: 1156–1166.
32. Dassa E, Bouige P. The ABC of ABCS: a phylogenetic and functional classification of ABC systems in living organisms. Res Microbiol 2001;152:211–229.
33. Fuellen G, Spitzer M, Cullen P, Lorkowski S. BLASTing proteomes, yielding phylogenies. In Silico Biol 2003;3:313–319.
34. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
35. Tomii K, Kanehisa M. A comparative analysis of ABC transporters in complete microbial genomes. Genome Res 1998;8:1048–1059.
36. Saurin W, Dassa E. Sequence relationships between integral inner membrane proteins of binding protein-dependent transport systems: evolution by recurrent gene duplications. Protein Sci 1994;3:325–344.
37. Tam R, Saier MH. Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. Microbiol Rev 1993;57:320–346.
38. Linton KJ, Higgins CF. The *Escherichia coli* ATP-binding cassette(ABC) proteins. Mol Microbiol 1998;28:5–13.
39. Braibant M, Gilot P, Content J. 2000. The ATP binding cassette (ABC) transport systems of *Mycobacterium tuberculosis*. FEMS Microbiol Rev 2000;24:449–467.
40. Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 1998;396:133–140.
41. Karlberg O, Canback B, Kurland CG, Andersson SG. The dual origin of the yeast mitochondrial proteome. Yeast 2000;17:170–187.
42. Gray MW, Burger G, Lang BF. The origin and early evolution of mitochondria. Genome Biol 2001;2:REVIEWS1018.
43. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA 2002;99:12246–12251.
44. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res 2002; 30:13–16.
45. Heringa J, Argos P. A method to recognize distant repeats in protein sequences. Proteins 1993;17:391–411.
46. Morgenstern B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 1999;15:211–218.
47. Brown NP, Leroy C, Sander C. MView: a web-compatible database search or multiple alignment viewer. Bioinformatics 1998;14:380–381.

48. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30:3059–3066.
49. Jonassen I. Discovering patterns conserved in sets of unaligned protein sequences. Methods Mol Biol 2000;143:33–52.
50. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–425.
51. Dean M. The human ATP-binding cassette (ABC) transporter superfamily. NCBI Bookshelf 2002. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.
52. Williamson P, Schlegel RA. Transbilayer phospholipid movement and the clearance of apoptotic cells. Biochim Biophys Acta 2002; 1585:53–63.
53. Madon J, Hagenbuch B, Landmann L, Meier PJ, Stieger B. Transport function and hepatocellular localization of mrp6 in rat liver. Mol Pharmacol 2000;57:634–641.
54. Gaedeke N, Klein M, Kolukisaoglu U, Forestier C, Müller A, Ansorge M, Becker D, Mamnun Y, Kuchler K, Schulz B, Mueller-Roeber B, Martinoia E. The *Arabidopsis thaliana* ABC transporter AtMRP5 controls root development and stomata movement. EMBO J 2001;20:1875–1887.
55. Lee EK, Kwon M, Ko JH, Yi H, Hwang MG, Chang S, Cho MH. Binding of sulfonylurea by AtMRP5, an *Arabidopsis* multidrug resistance-related protein that functions in salt tolerance. Plant Physiol 2004;134:528–538.
56. Forestier C, Frangne N, Eggmann T, Klein M. Differential sensitivity of plant and yeast MRP(ABCC)-mediated organic anion transport processes towards sulfonylureas. FEBS Lett 2003;554: 23–29.
57. Blair JE, Ikeo K, Gojobori T, Hedges SB. The evolutionary position of nematodes. BMC Evol Biol 2002;2:7.
58. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivares CP. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 2001;414: 450–453.
59. Perrière G, Gouy M. WWW-query: an on-line retrieval system for biological sequence banks. Biochimie 1996;78:364–369.
60. Bisbal C, Martinand C, Silhol M, Lebleu B, Salehzada T. Cloning and characterization of a RNAse L inhibitor: a new component of the interferon-regulated 2-5A pathway. J Biol Chem 1995;270: 13308–13317.
61. Zillig W, Klenk HP, Palm P, Leffers H, Pühler G, Gropp F, Garrett RA. 1989. Did eukaryotes originate by a fusion event? Endocytobiosis Cell Res 1989;6:1–25.
62. Sanchez-Fernandez R, Davies TG, Coleman JO, Rea PA. The *Arabidopsis thaliana* ABC protein superfamily, a complete inventory. J Biol Chem 2001;276:30231–30244.
63. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–29.
64. Pérez AJ, Rodríguez A, Trelles O, Thode G. A computational strategy for protein function assignment which addresses the multidomain problem. Comp Funct Gen 2002;3:423–440.
65. Bellosta S, Bernini F, Chinetti G, Cignarella A, Cullen P, von Eckardstein A, Exley A, Freeth J, Goddard M, Hofker M, Kanters E, Kovanen P, Lorkowski S, Pentikainen M, Printen J, Rauterberg J, Ritchie A, Staels B, Weitkamp B, de Winther M. Macrophage Function and Stability of Atherosclerotic Plaque Consortium: Macrophage function and stability of the atherosclerotic plaque: progress report of a European project. Nutr Metab Cardiovasc Dis 2002;12:3–11.