

# ABCdb: an ABC Transporter Database

Yves Quentin\* and Gwennaele Fichant

Laboratoire de Chimie Bactérienne, Institut de Biologie Structurale et Microbiologie, CNRS, 31, Chemin Joseph Aiguier, 13402, Marseille Cedex 20, France

## Abstract

**We present the first release of a database devoted to the ATP-binding cassette (ABC) protein domains (ABCdb). The ABC proteins are involved in a wide variety of physiological processes in Archea, Bacteria and Eucaryota where they are encoded by large families of paralogous genes. The majority of ABC domains energize the transport of compounds across the membranes. In bacteria, ABC transporters are involved in the uptake of a wide range of molecules and in mechanisms of virulence and antibiotic resistance. In eukaryotes, most of them are involved in drug resistance and in human cells, many are associated with diseases. Sequence analysis reveals that members of the ABC superfamily can be organized into sub-families and suggests that they have diverged from common ancestral forms. In this release, ABCdb includes the inventory and assembly of the ABC transporter systems of completely sequenced genomes. In addition to the protein entries, the database comprises information on functional domains, sequence motifs, predicted trans-membrane segments, and signal peptides. It also includes a classification in sub-families of the ABC systems as well as a classification of the different partners of the systems. Evolutionary trees and specific sequence patterns are provided for each sub-family. The database is endowed with a powerful query system and it was interfaced with blastP2 program for similarity searches. ABCdb has been developed in the ACeDB format, a database system developed by Jean Thierry-Mieg and Richard Durbin. ABCdb can be accessed via the World Wide Web (<http://ir2lcb.cnrs-mrs.fr/ABCdb/>).**

## Introduction

ABC transporter systems, also termed traffic ATPases, are found in the three major kingdoms of life (Prokaryota, Archea and Eukaryota, reviewed by Higgins, 1992). The majority mediate the active uptake or efflux of specific molecules across the biological membranes. They handle a wide variety of compounds, which differ in nature and size (*i.e.* oligosaccharides, amino acids, peptides, antibiotics, metallic cations... Ames, 1986). A typical ABC transporter is composed of two Membrane Spanning Domains (MSD) and two Nucleotide Binding Domains

(NBD). The import systems are associated with a Solute Binding Protein (SBP). The MSDs constitute the membrane channel and the NBDs, in close interaction with the MSDs, energize the transport via ATP hydrolysis. The SBPs are soluble and periplasmic in Gram negative bacteria and anchored to the membrane in Gram positive bacteria. The SPBs confer specificity for compounds to the transporter. In bacteria, ABC systems are generally encoded by neighboring genes. In eukaryotes, they usually correspond to a single amino acid chain and only export systems have been described. Many of them are involved in multi-drug resistance or genetic diseases (for review, see Holland and Blight, 1999).

Recently, the question of the inventory of the ABC transporters in yeast and bacterial complete sequenced genomes has been addressed by several authors (Taglicht and Michaelis, 1997; Decottignies and Goffeau, 1997; Paulsen *et al.*, 1998a, 1998b; Linton and Higgins, 1998; Quentin *et al.*, 1999; Dassa *et al.*, 1999; Tomii and Kanehisa, 1998), and these studies revealed that ABC systems can be arranged in a comprehensive classification that is well correlated with compound specificity of transport. Comparative analyses of repertory of ABC transporters between genomes suggest that ABC transporters derive from successive waves of duplications and that the ancestral ABC transporter may have arisen early in evolution, before the differentiation of prokaryotes and eukaryotes, in the last common universal ancestor (Tomii and Kanehisa, 1998 and Saurin *et al.*, 1999). These observations motivate the development of a database dedicated to ABC transporters. Such a database should be useful to predict the compound specificity of hypothetical transporters obtained through systematic sequencing, but should be also relevant for studies of multi-drug resistance of eukaryotic cells, virulence and antibiotic resistance of prokaryotes. This database will constitute a very good platform for further evolutionary and structure-function relationship studies.

## Source of Data

The primary data are retrieved from the Genome page of the NCBI server (<http://www3.ncbi.nlm.nih.gov/Entrez/Genome/org.html>) as GenBank flat files. The files are reformatted with perl programs before their importation entries in ABCdb. In the public version, only peptides and information on proteins related to ABC transporters are available. As far as possible, the protein names are those given in the GenBank entries to allow cross-linking between NCBI and ABCdb. The strategies we have adopted for automatic partner recognition is summarized on our WEB site (<http://ir2lcb.cnrs-mrs.fr/ABCdb/>) and will be published elsewhere.

## Database Design

We administer our data with the ACeDB system developed by Thierry-Mieg and Durbin (ACeDB is an acronym for A C*aenorhabdis* e*legans* DataBase). ACeDB was created

Received May 12, 2000; revised May 23, 2000; accepted May 23, 2000.  
\*For correspondence. Email [quentin@ir2lcb.cnrs-mrs.fr](mailto:quentin@ir2lcb.cnrs-mrs.fr); Tel. 33-4 91 16 44 12; Fax. 33-4 91 71 89 14.



## Model: ?Assembly

?Assembly	Title	UNIQUE	?Text
	Species	UNIQUE	?Species XREF Assembly
	Remark	?Text	
	Partners	?Protein	Text
	Domain_Organization	UNIQUE	?Text
	Functional_Classification	UNIQUE	?Classif XREF Member
	Function	?Text	

For technical information about these pages see:

[AceDB Home Page](#)

[AcePerl Home Page](#)



## Assembly: Ecol.ARA

Ecol.ARA	Species	Escherichia coli
	Partners	Ecol.ARAG_N1 NBD1 Ecol.ARAG_N2 NBD2 Ecol.ARAH_1 MSD1 Ecol.ARAH_2 MSD2 Ecol.ARAF SBP
	Domain_Organization	NBD1-NBD2,MSD1,MSD2,SBP
	Functional_Classification	A_1a
	Function	arabinose uptake

Figure 1. Composite figure with the display of the model of the class Assembly and of the assembly object Ecol.ara. The icon "Search" links to the query tools, "Graphic Display" allows a graphical representation of the object (useful for tree and protein objects), and "Tree Display" is the current display. The tag names are in bold and the texts (underlined) are links to other classes. Tag with the keyword UNIQUE cannot be repeated in objects when the other can have several entries (*i.e.* an assembly object may contain more than one protein partner). The tag XREF is used to force cross-referencing between two classes. The assembly Ecol.ara is composed of one dimeric ATPase (Ecol.ARAG), two membrane proteins (Ecol.ARAH\_1 and Ecol.ARAH\_2), and one solute binding protein (Ecol.ARAF). The transporter belongs to functional class A\_1a and is involved in the arabinose uptake.

for managing the nematode genome project and in this context the system and the database have the same name. In the ACeDB system the data are stored in objects, which fall in a number of classes and each class is defined by a model. In the implementation of the ABCdb, we created some new classes "Assembly", "Classif", "Sub-family", "Pattern" and "Profile". The class Assembly is used to describe multi-protein systems such as ABC transporters (Figure 1). Its model includes tags for the origin, the (putative) function of the system and for the protein partners involved in the multi-protein assembly. For ABC systems, the partners are decomposed as functional domains (NBD, MSD, and SBP) and a tag (Domain\_Organization) is used to define the contribution of each domain in the system. As illustration, the ABC transporter of the arabinose uptake is given Figure 1. Since the multi-protein assembly is the functional unit, we associated the functional classification of the ABC transporters to the "Assembly" class. The functional classification is modeled by a class "Classif" that is fairly simple in this primary version of the database. The purpose of this class is to collect and summarize the experimental data available on the systems. This work cannot be done automatically and will be achieved progressively by manual contributions of human experts.

On the other hand, the proteins and proteins domains can be automatically classified in sub-families with computer tools based on combinations of profiles (the pattern). Therefore, we design three new classes corresponding to the sub-families, the patterns and the profiles. Each sub-family is defined by a pattern, which includes one or several profile(s) computed by the MEME program (Bailey and Elkan, 1994).

The model of the "Protein" class has been fairly

simplified regarding the ACeDB distribution package but few lines have been added for the creation of links with the new classes. The public version of the database does not include DNA sequence data but the coordinates of the genes on the chromosome are included in the "Protein" class. On the WEB version, hyperlinks to the NCBI genomic display are automatically generated.

We used the recent development of dendrogram tree display due to Richard Bruskiwich (Sanger Centre) to represent trees. Two kinds of trees are available in the database: a tree reflecting the taxonomy of the genomes as it can be found on the NCBI server, and evolutionary trees obtained on the proteins of each sub-family. Each evolutionary tree is computed with the NJ method (Saitou and Nei, 1987) from a distance matrix based on a multiple alignment obtained with ClustalW (Thompson *et al.*, 1994). However, such trees, automatically generated, should be considered only as indicative (Figure 2). One can navigate along the tree and display the information associated with the nodes or the leaves with the mouse button. This is a simple way to browse through the information contained in the database.

Relationships between the classes of the database core are summarized in Figure 3. The Protein class has a central position with links to all the other facets of the ABCdb: the systems (Assembly and Classif), the motifs (Profile and Pattern), the sub-families (Sub-family and Family), the genomes (Species and Taxon), and the trees (Tree and Treenode). Since each pattern is diagnostic of a sub-family, there is a direct link between both classes. The link to Species is added in Assembly model to facilitate the cross-queries between these classes.



**Tree: tree\_S1a** *Requires IE or Netscape version 4.0 or higher.*

# Leaves: 15    Max. Branch Length: 2.211  
 Display Normalization: 40.0

Scale: 4.0 = 0.221

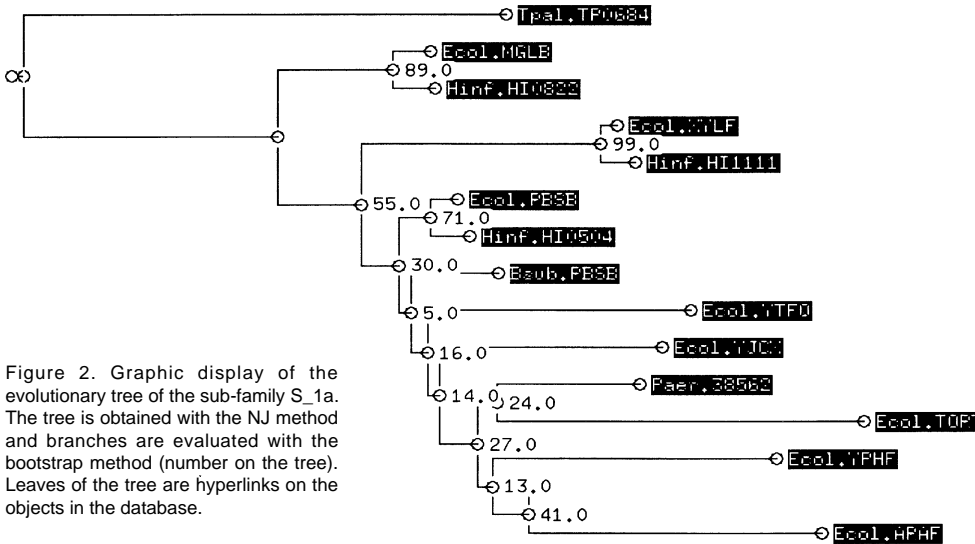


Figure 2. Graphic display of the evolutionary tree of the sub-family S\_1a. The tree is obtained with the NJ method and branches are evaluated with the bootstrap method (number on the tree). Leaves of the tree are hyperlinks on the objects in the database.

**Data Query**

The simplest way to explore the database is to follow the links present in the objects, or to search for objects in a given class or in all classes (See Figure 4, "Simple Search" option). The search can be restrict by a sub-string included in the name of the objects to retrieve. The result appears as a list of object names and each object can be displayed with a mouse "double-click" on its name.

The ACeDB system also includes a powerful query language (See Figure 4, "Ace Query" option). Since, the queries applied to tag and data contained in objects, the knowledge of the class models is required. Therefore, the first query could be to list the database models (enter the query: "find model"). A query is composed of commands, operators and patterns. The find command is used to retrieve the list of objects belonging to a given class and matching the supplied pattern. Composite queries can be achieved by chaining a series of simple queries separated by semicolons. Each query is applied to the list of objects retrieved by the previous one. The follow command can be used to retrieve objects attached to a specified tag present in a list of objects selected by a find command.

As examples of find command, the simple queries "find Protein Ecol\*" lists all proteins with names that begin with Ecol\* and "find Protein Membrane\_segment < 5 OR Membrane\_segment > 8" gives all proteins containing less than 5 TMs or more than 8 TMs. As example of complex queries using the follow command, "find Species Lineage = Archaea; follow Assembly; follow Functional\_Classification" gives the list of ABC system sub-families present in Archaea, and "find Assembly Domain\_Organization = \*SBP\*; follow Partners; Domain\_Structure = \*MSD\* AND NOT Profil\_homol = EAA"

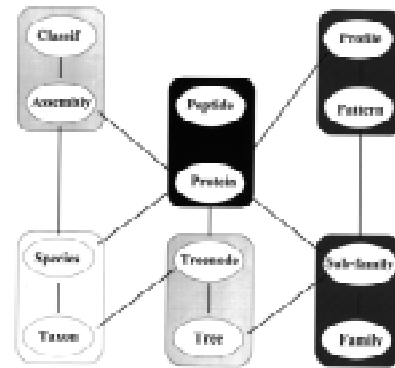



Figure 3. Relationships between the major classes of the database.

retrieves the membrane proteins of uptake systems that do not contain the EAA motif. Detailed instructions and other queries can be found at <http://ir2lcb.cnrs-mrs.fr/ABCdb/>.

An another powerful tool is the "TableMaker", which does not retrieve lists of objects but builds complex tables from data gathered in the same or different classes. Each column in the table is the result of a query. One application of the "TableMaker" is to tabulate the partners of each ABC system according to a criterion such as genome origin or sub-family classification. This tool will not be further described, as it is not yet available in the web version of the database, but example of tables are provided on our web site.

<a href="#">Simple Search</a>	<a href="#">Text Search</a>	<a href="#">Class Browser</a>	<a href="#">Ace Query</a>	<a href="#">Blast Search</a>
-------------------------------	-----------------------------	-------------------------------	---------------------------	------------------------------

 **Simple Search**

Select the type of object you are looking for and optionally type in a name or a wildcard pattern (? for any one character. \* for zero or more characters). If no name is entered, the search displays all objects of the selected type. *Anything* searches for the entered text across the entire database.

( Anything ( Protein ( Tree **Name:**

( Specie ( System

Figure 4: Entry page of ABCdb showing the menu bar and a simple query on the ABC transporter for the arabinose in *Escherichia coli*. The "Simple Search" option is used to retrieve objects by name among a selection of classes. The "Text Search" option looks for string patterns in object names (fast search) or in all data entries (long search). The "Class Browser" option allows the selection of objects by class. The "Ace Query" option is used to compose complex queries. The "Blast Search" option is a blastP2 interface for the database.

## Protein Analysis Tools

We implemented a blastP2 form on our web server (Figure 5). The user query is compared to all proteins of ABCdb and html links to the protein objects and sub-families are added in the output in order to facilitate the classification and comparison of the query with ABCdb entries.

## ACeDB Software and Data Access

The public version of ABCdb is available at the following address, <http://ir2lcb.cnrs-mrs.fr/ABCdb/>. Comments, suggestions and corrections can be addressed to [quentin@ir2lcb.cnrs-mrs.fr](mailto:quentin@ir2lcb.cnrs-mrs.fr). Users of ABCdb are politely asked to cite this article within scientific publications related to its use.

## Acknowledgements

We thank Elie Dassa and François Denizot for advice and encouragement, Athel Cornish-Bowden for critical reading of the manuscript and Abdelkader Berkane, Jean Christophe Perennes and Philippe Mateos for their contribution in the annotation of ABCdb. We are grateful to Jean Thierry-Mieg (CNRS, Montpellier) and Richard Durbin (Sanger Centre) for the ACeDB system, Lincoln D. Stein (Cold Spring Harbor Laboratory) for the AcePerl (Web interface for ACeDB), and Richard Bruskiwich (Sanger Centre) for the development of dendrogram tree. This work was supported by the CNRS (Centre National de la Recherche Scientifique) and the ARC (Association pour la Recherche sur le Cancer) grant 5268.

## References

Ames, G.F. 1986. Bacterial periplasmic transport systems: structure, mechanism, and evolution. *Annu. Rev. Biochem.* 55: 397-425.

Bailey, T.L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *ISMB* 2: 28-36.

Dassa, E., Hofnung, M., Paulsen, I.T., and Saier, M.H. Jr 1999. The

*Escherichia coli* ABC transporters: an update. *Mol. Microbiol.* 32: 887-889.

Decottignies, A., and Goffeau, A. 1997. Complete inventory of the yeast ABC proteins. *Nature Genet.* 15: 137-145.

Higgins, C.F. 1992. ABC transporters: from microorganisms to man. *Annu. Rev. Cell Biol.* 8: 67-113.

Holland, B., and Blight, M., A. 1999. ABC-ATPases, adaptable energy generators fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans. *J. Mol. Biol.* 293: 381-399.

Linton, K.J., and Higgins, C.F. 1998. The *Escherichia coli* ATP-binding cassette (ABC) proteins. *Mol. Microbiol.* 28: 5-13.

Paulsen, I.T., Sliwinski, M.K., Nelissen, B., Goffeau, A., and Saier, M.H. Jr 1998a. Unified inventory of established and putative transporters encoded within the complete genome of *Saccharomyces cerevisiae*. *FEBS Lett.* 430: 116-125.

Paulsen, I.T., Sliwinski, M.K., and Saier, M.H. Jr 1998b. Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.* 277: 573-92.

Quentin, Y., Fichant, G., and Denizot, F. 1999. Inventory, assembly, and analysis of *Bacillus subtilis* ABC transport systems. *J. Mol. Biol.* 287: 467-484.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406-425.

Saurin, W., Hofnung, M., and Dassa, E. 1999. Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. *J. Mol. Evol.* 48: 22-41.

Taglicht, D., and Michaelis, S. 1998. A complete catalogue of *Saccharomyces cerevisiae* ABC proteins and their relevance to human health and disease. *Meth. Enzymol.* 292: 130-162.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.

Tomii, K. and Kanehisa, M. 1998. A comparative analysis of ABC transporters in complete microbial genomes. *Genome Res.* 8: 1048-1059.

Walker, J.E., Saraste, M., Runswick, M.J., and Gay, N.J. 1982. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* 1: 945-951.

<a href="#">Simple Search</a>	<a href="#">Text Search</a>	<a href="#">Class Browser</a>	<a href="#">Ace Query</a>	<a href="#">Blast Search</a>
-------------------------------	-----------------------------	-------------------------------	---------------------------	------------------------------

**abcace BLAST Results**

**Results Summary**

Sequence	SF	Dom	Description	Score	E Value
<a href="#">Ecol.ARAF</a>	S_1a	SBP	L-arabinose-binding periplasm...	637	0.0
<a href="#">Paer_38562</a>	S_1a	SBP	D-ribose transport*	60	1e-10
<a href="#">Ecol.MGLB</a>	S_1a	SBP	galactose-binding transport p...	56	2e-09
<a href="#">Ecol.RBSE</a>	S_1a	SBP	D-ribose periplasmic binding ...	52	2e-08
<a href="#">Hinf.H10822</a>	S_1a	SBP	galactose ABC transporter, ...	46	2e-06
<a href="#">Hinf.H10504</a>	S_1a	SBP	D-ribose ABC transporter, p...	46	2e-06
<a href="#">Ecol.Y1FQ</a>	S_1a	SBP	putative LACI-type transcript...	42	3e-05
<a href="#">Rspe.Y4MI</a>	S_1a	SBP	Y4mi*	37	0.001

8 hits total (8 shown)

Figure 5: BlastP2 results for Ecol.ARAF query. The output is fairly classical but we have added links to proteins and sub-families in ABCdb and information on the domain organization of the proteins. The sequence alignments have been removed.