CRANFIELD UNIVERSITY

M NAVAL SÁNCHEZ

PHYLOGENETIC ANALYSIS BLACK BOX

CRANFIELD HEALTH

MSc THESIS

CRANFIELD UNIVERSITY

CRANFIELD HEALTH

MSc THESIS

Academic Year 2007-2008

M NAVAL SÁNCHEZ

Phylogenetic Analysis Black Box

Supervisors: Dr J. Huxely-Jones (GSK) & Dr L. Larcombe

September 2008

This thesis is submitted in partial fulfilment of the requirements
for the degree of Master of Science

Abstract

Molecular phylogenetics tries to resolve sequence evolution in order to provide the evolutionary relationship of homologous sequences: DNA, RNA or proteins. Phylogenetics acts at the beginning of the drug discovery process and is of great relevance in gene target selection, protein functionality inference and animal model selection. There are different methodologies to develop a phylogenetic analysis. However, none of them are 100% accurate. Therefore, it is not correct to rely on only one phylogenetic methodology to arrive at significant conclusions. The ideal use of phylogenetics is to perform a "robust" phylogenetic analysis that would involve the use of various methodologies and a comparison among their results. As most bioinformaticians performing phylogenetic analysis have expert knowledge in phylogenetics, this project intends to be a phylogenetic guide. Firstly, it provides a detailed introduction of phylogenetics giving a description of the most relevant methodologies. Secondly, a phylogenetic analysis of the superfamily ABC transporters was performed to provide an example of how to perform robust manual phylogenetics analysis and interpret these results. Furthermore, because of the relevance in pharmacology of the ABC transporters, discussion of the results here provided can be used to help the understanding of their ABC transporters evolution and implications in disease. Finally, the Phylogenetic analysis Black Box tool (PBB) has been generated. The program is a pipeline of the most relevant phylogenetic methods and provides easy and fast access to the non-phylogenetic expert user, allowing them to perform a robust phylogenetic analysis, in a time saving way and helping them in the results interpretation.

Acknowledgements

First and foremost I would like to thank my supervisor, Dr Julie Huxley-Jones, for relying on me throughout this project, and for all the attention and patience provided. Thanks for your support, guidance and corrections. Thanks to my supervisor, Dr Lee Larcombe, for his advice and guidance.

I would like to thank everyone at the Computational Biology Department at GSK in Harlow who encouraged me in the hard times, especially to my colleague Cathy Mitchell. I am also extremely grateful to Steve Deharo, for his programming advice, and to Samuil Hasan and Mark Simmons for their help in IT issues. Thanks also to Aaron J. Mackey for his brilliant ideas to improve my script. I would like to thank everyone in the department that relied on the program and helped me in its testing, in special to Simmon Topp and Steve Deharo.

Thanks to all the people I met this year and have made me enjoy my stay at Cranfield, classmates Barnabas, Rui and Tim and to Amit and Fady who apart from helping me in bioinformatic problems have treated me as a sister.

This project would have not been feasible without the support of my family. Thanks to them for their love, their trust and their comprehension. A huge thank as well to my friend Cristina for being always there and affective thank to you, Jeremy.

Contents

List of Tables

List of Figures

Abbreviations

ABC	ATP binding- cassete
ALD	Adrenoleukodystrophy
ATP	Adenosine Triphosphate
BLAST	Basic Local Alignment Search Tool
BLOSUM	BLOcks of Amino Acid SUbstitution Matrix
CFTR	Cystic Fibrosis Transmembrane
DNA	Deoxyribonucleic Acid
IBM	International Business Machines
GSK	GlaxoSmithKline Pharmaceuticals
LCA	Last Common Ancestor
MCMC	Markov Chain Monte Carlo
MDR	Multidrug Resistance
MRP	Multidrug Resistance Proteins
MultiAlin	Multiple sequence Alignment
MUSCLE	Multiple Sequence Comparison by Log-Expectation
NBF	Nucleotide-Binding Fold
NCBI	National Center of Biotechnology Information
OABP	Oligo-adeylate-binding protein
OMIM	Online Mendelian Inheritance in Man
OTUs	Operational Taxonomic Units
PAM	Point of Accepted Mutation
PAUP	Phylogenetic Analysis Using Parsimony
PBB	Phylogenetic analysis Black Box
Perl	Practical Extraction and Reporting Language
PHYLIP	Phylogeny Inference Package
ProML	Protein Maximum Likelihood program
Protdist	Program to compute distance matrix from protein sequences
Protpars	Protein sequence parsimony method
RACE	Rapid Amplification of cDNA Ends
RNA	Ribonucleic acid
Seqboot	Sequence bootstrap
TAP	Antigenic Peptide Transporter
T-Coffee	Tree based Consistency Objective Function For AlignmEnt Evaluation
TM	Transmembrane domain
VAST	Visualisation of Aligned Sequences and Trees

Chapter 1. Introduction

1.1 Phylogenetics

Phylogenetics is the study of the evolutionary relationships of the living organisms being represented by treelike diagrams. The term phylogenetics is derived from phylogeny, a word defined in 1866 by Erns Haeckel as “the evolutionary history of life” (Hillis, 1997a). However at the origin it referred to any morphological classification, nowadays the data used to develop phylogenetic analysis are morphological, behavioural or molecular. The latter is the most recent kind of data used in phylogenetic analysis and the one that has become of great interest in bioinformatics.

Molecular phylogenetics appeared in the 1960s as a new area in biology. Its emergence reflected the amount of data provided by new molecular biology techniques, such as immunological analysis, protein electrophoresis or DNA-DNA hybridisation, and became common use in the scientific community in the 1980s (Hillis, 1997b). In 2003, after the Human Genome Project a “boom” in phylogenetics occurred due to the arrival of robust and fast sequencing methods. Molecular phylogenetics is now defined, as “the study of evolutionary relationships of sequences of nucleotides or proteins based on the mutations at various positions in the sequences” (Xiong, 2006).

Molecular phylogenetics can serve to depict the evolutionary relationship of different living organisms. Further, molecular phylogenetics can be used to illustrate the evolutionary relationship of macromolecular entities (sequences of nucleotides or proteins), to show the evolution of a gene family, or of genes or proteins that depict the same function.

1.2 Homology

Homology can be defined as the amount of similarity between sequences (Baldauf, 2003). Therefore, homologues are after sequences that share common ancestry. The phylogenetic relationships between different sequences are based on homology. There are three homologue types (Baxevanis and Ouellette, 2004; Claverie and Notredame, 2007):

- **Orthologue:** homologue produced by speciation.
- **Paralogue:** homologue due to gene duplication.
- **Xenologues:** homologue originated by horizontal gene transfer between two organisms.

1.3 Homologue identification

In order to know if two sequences are homologues, bioinformatics relies on similarity, where a pair of sequences are considered homologues, and difference between them a consequence of divergence in evolution. There is no method that confirms that two sequences share common ancestry. However, there are algorithms that search for similarity between sequences. These algorithms are known as alignments. The most used alignment to perform holomologue identification is BLAST (Basic Local Alignment Search Tool) (Altschul *et al.*, 1997; Schäffer *et al.*, 2001).

1.4 Alignments

Alignments are algorithms to compare the similarity of sequences: nucleotides or amino acids. Their main goal is to put the sites of the sequences to be compared in the order that generates the maximum number of site matches between them. If the sequences to be aligned are homologues, the matches between sites will reveal the positions that have been conserved in evolution and the mismatches the positions that have diverged. There are two kinds of alignments depending on the number of sequences to compare: pairwise alignment and multiple alignment.

1.4.1 Pairwise Alignment

Pairwise alignment algorithms are built to compare just two sequences. In phylogenetics they are used to infer the homology of the sequences compared. The most commonly used algorithms are discussed below.

The Needleman and Wunsch (Needleman and Wunsch, 1970) algorithm generates a global alignment, and compares sequences in their entire length using a two-dimension matrix (MAT). Inside the matrix there are values containing information about the probability of the change from Sequence A to Sequence B. The Maximum-match pathway is the path that starting from the last position of MAT towards the origin searching provides the highest score.

The Smith and Waterman (Smith and Waterman, 1981) algorithm is a modification of the Needleman and Wunsch algorithm in order to provide a local alignment, when segments or regions with high similarity are searched. The algorithm follows the same process as Needleman and Wuncsh alignment but allows the possibility of ending the segments at any position of the two sequences, has differences in gap penalties, and the the starting point is the position with highest score.

Due to the computational load required to calculate the previous alignments, heuristic methods were designed. BLAST (Basic Local Alignment Search Tool) (Altschul *et al.*, 1997; Schäffer *et al.*, 2001), is the most commonly used heuristic method. It allows comparing sequence similarity against all the sequences of a database in a time saving mode. The first BLAST algorithm was reported in 1990 (Altschul *et al.*, 1990) and has been improved in 1997 (Altschul *et al.*, 1997). The main feature of BLAST is that instead of considering all the sequence in order to perform the alignment, it splits the sequences in subunits called words (usually length of three amino acids for proteins). The similarity search of a sequence is performed aligning its words against the sequences from different databases of interest. If the score of the word against one sequence is greater than *T* (BLAST parameter known as threshold) the word is

called a “hit” and the algorithm expands it on both sides. The hit expansion continues until its score starts decreasing. The improved algorithm changes the requirements and only considers alignments where at least two hits are reported. The alignments with highest score would be the ones reported to the user. However, the amount of alignments to perform is bigger than considering the entire sequence, word lengths are much shorter, therefore the alignments to carry out are much faster. Currently, BLAST is a family of sequence query specific programs (Table 1). The user can choose one or another depending on their interest. The specification of each program improves the search accuracy.

Table 1. List of BLAST family programs provided by (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008).

BLAST family programs	
nucleotide blast	Search a nucleotide database using a nucleotide query
protein blast	Search protein database using a protein query
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

1.4.2 Multiple sequence alignment

Multiple sequence alignments are significantly more relevant in phylogenetics, as they offer the alignments of multiple taxa to infer phylogeny from. In multiple sequence alignment all the existent algorithms are heuristic. Exhaustive algorithms are not feasible even when the user has as few as than four sequences to align (Baxevanis and Ouellette, 2004). The most representative kinds of multiple sequence alignments are presented as follows:

The most used and accurate method to perform multiple sequence alignment is progressive alignment (Baxevanis and Ouellette, 2004). Progressive alignments are based on Needleman and Wunsch pairwise analysis for each pair of sequences in the multiple alignment. The score results are stored in a distance

matrix. A basic guide phylogenetic tree is also executed. Once the guide tree is produced, the sequences would be aligned following in an iterative process:

1. Two sequences are aligned by pairwise alignment.
2. A consensus sequence is generated.
3. Alignment of the consensus sequence with the most closely related sequence according to the guide tree.
4. A consensus sequence of the three sequences is generated.
5. Repetition of steps 3 and 4 until all the input sequences are aligned.

Commonly used multiple sequence alignment software performing progressive alignments are: CLUSTALX (Thompson *et al.*, 1997), CLUSTALW (Thompson *et al.*, 1994), T-Coffee (Notredame *et al.*, 2000) and MUSCLE (Edgar, 2004a; Edgar, 2004b).

The fundamentals of the iterative alignment method are to align the sequences randomly and keep on improving the alignment by several realignments (Chakrabarti *et al.*, 2006). Iterative alignment is mostly used to refine alignments provided by the progressive alignment algorithm. If it is used as a multiple sequence alignment alone there is no guarantee that the optimal solution has been achieved (Notredame and Higgins, 1996). Multiple sequence alignment software performing Iterative alignments include PRRN (Gotoh, 2007) and MultiAlin (Corpet, 1988).

Block-based alignment methods run a local alignment. This methodology is utilised when there is only local similarity among the sequences, for example, domains or motifs (Lassmann and Sonnhammer, 2002), and is implemented in software such as DIALIGN (Subramanian *et al.*, 2005).

1.5 Models of Evolution

The divergence of two sequences can be calculated by the number of substitutions in an alignment. However, not all substitutions in a nucleotide or amino acid sequence occur at the same frequency. Models of evolution or substitution models are the criteria followed to provide scoring matrices depending on the likelihood of replacement of for every pair of residues in a sequence. Nucleotide substitution matrices are composed of four nucleotides Adenine (A), Timine (T), Cytosine (C), Guanine (G). The frequencies of mutation are not equal for all pairs of bases and substitution matrices give different scores to transitions and transversions (Kimura, 1980). In amino acid substitution matrices there are hundreds of possible amino acid substitutions (21 x 21) to score. Amino acid mutations are most probable between amino acids sharing similar properties, minimising protein structure and functionality changes. There are also scoring matrices that obtain the replacement scores considering physicochemical properties of the amino acids and the genetic code interchangeability. However, methods that calculate their scores empiric analysis are considered more reliable. PAM and BLOSUM are the most extended amino acid substitution matrices.

1.5.1 Point of Accepted Mutation matrices (PAM)

The PAM method works on the principle “replacement of one amino acid by another, accepted by natural selection” (Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978). One PAM unit corresponds to a unit of evolutionary divergence in which 1% of the amino acid have been changed. PAM matrices are used in phylogenetics because their scores involve evolutionary adaptation. However, they do not have good performance when divergent sequences are compared. Matrices following the same approach as PAM but using bigger datasets (Gonnet matrices (Gonnet *et al.*, 1992) and Jones-Taylor-Thornton matrices (Jones *et al.*, 1992)) have proved reliable in phylogenetic tree construction (Xiong, 2006).

1.5.2 BLOcks of Amino Acid SUbstitution Matrix (BLOSUM)

The BLOSUM method (Henikoff and Henikoff, 1992) is based on clustering different protein blocks[1] with a predetermined identity level of the segments[2] to compare. From that data set the odd ratios for an amino acid to be substituted by another are calculated. The final scores show the amino acid substitution rates as the amount of frequencies a substitution has been observed by the number of times expected to occur by chance, different BLOSUM matrices are available depending on the identity level chosen from the segments to be compared. BLOSUM 62 and BLOSUM 80 are the most commonly used. BLOSUM matrices are not based in any evolutionary method. Therefore, they perform better when aligning less closely related sequences.

1.6 Phylogenetic Tree Building Methods

Phylogenetic tree building methods can be described as the algorithms whose function is to infer phylogenies from the taxa to study. There is more than one method due to each of them are based in a distinct but valid mode to infer evolutionary relationships. They can be classified into parametric, they use a determined model of evolution, or non-parametric, based on a determined metric to generate the tree and non model of evolution is considered. The parametric methods embrace the distance methods and Maximum Likelihood methods. A commonly used non parametric method is Maximum Parsimony.

1.6.1 Parametric methods

1.6.1.1 Distance methods

Distance methods infer phylogenetic trees based on the genetic distances between pairs of sequences. Genetic distances are the dissimilarity between pairs of sequences, where an amino acid has been changed (Xiong, 2006). Once the distances are calculated the character data is no more taken into account.

[1] Block: conserved region of a protein family. The blocks where extracted from PROTOMA database (Henikoff and Henikoff, 1992).
[2] Segment: part of a protein sequence of width equal to the block (Henikoff and Henikoff, 1992).

Neighbor Joining (Saitou and Nei, 1987) is the most commonly used distance based method (Bryant, 2005). The procedure of this methodology is referred as *star decomposition*. Initially, all the taxons are considered in the same cluster, the starting phylogenetic tree is a starlike tree (Figure 1a). Once the clustering of the most similar taxons is resolved two clusters are obtained (Figure 1b) (star decomposition). This process will be repeated for the reformed star tree. The process will end once all the taxons are clustered. It can be that at the first stage (star tree like) there are more than one pair equally distanciated, and so, Neighbor joining methods produce several subtrees choosing once the process has ended the most suitable for the analysed data. It is remarkable to say that Neighbor Joining method generates unrooted trees and do not consider molecular clock, homogeneous rates of evolution among sites. Neighbor Joining does not consider that more than one mutation has occurred in one site, "mutation saturated sequence".

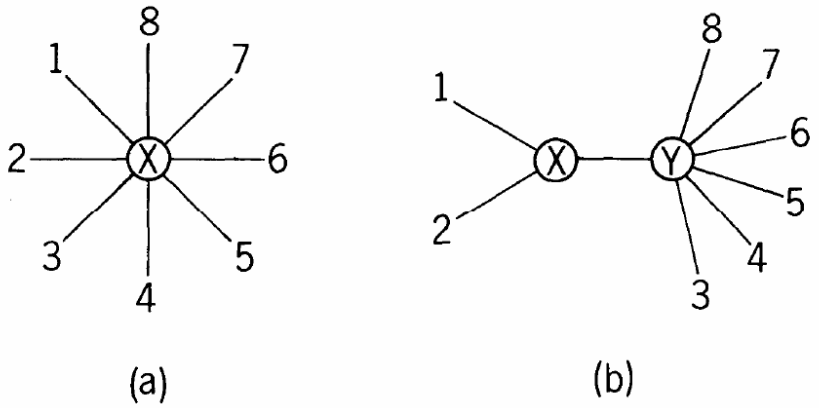


Figure 1. a) The first starlike tree, no hierarchical structure. b) Stepwise Neighbor Joining method. Taxons 1 and 2 have been clustered because the have the smallest sum of branch lengths. Figure extracted from Saitou and Nei (1987).

The most frequently used software for performing Neighbor Joining is PHYLIP with the applications Protdist and Neighbor (Felsenstein, 2005). Protdist computes the distance matrix of the original dataset based in the evolutionary model selected. Neighbor infers the phylogeny following the Neighbor Joining method, star decomposition, according to the distances provided by Protdist.

1.6.1.2 Maximum Likelihood methods

Maximum Likelihood algorithms refer to different character based methodologies that base their results in statistical models. Here the most extended maximum algorithms, maximum likelihood and bayesian, are presented.

Maximum Likelihood methodology provides the most likely tree from a given dataset prior to determination of an evolution model (Mount, 2004). The first step of the algorithm is to generate all the possible rooted topologies for the number of sequences to study. Then, starting by one topology a random order of the amino acids located on the first site of the sequences are put in the outer leaves of the selected topology. All possible inner nodes that will give rise to the outer sequences are generated. One of the patterns is put in the tree. The likelihood of topology + site amino acid order + pattern inner node is calculated based on the model of evolution chosen (cf. Section 1.5). This method is repeated for each pattern in the inner node, amino acid site order and possible topology and for each site. As a consequence the maximum likelihood method is highly computationally demanding. The most likely overall tree is the one that gives the highest overall probability at all the sites found by summing the site probabilities for each tree (Mount, 2004). Softwares performing Maximum Likelihood algorithm are PAML (Yang, 1997; Yang, 2007), PHYLIP package within the ProML application (Felsenstein, 2005) and Tree Puzzle (Schmidt *et al.*, 2002).

The maximum likelihood algorithm can also input the rate of variation of the amino acid sites. The rate of heterogeneity among sites is considered in most of the cases to follow a gamma distribution curve, in a combined approach of invariable sites and variable sites following the gamma distribution (Baxevanis and Ouellette, 2004). Softwares performing the gamma distribution generation include Tree Puzzle (Schmidt *et al.*, 2002).

Bayesian analysis is considered as a maximum likelihood methodology, but its fundamentals are based on Bayes theorem (Eq. 1)

Equation 1 (Huelsenbeck et al., 2001):

$$\Pr[Tree \mid Data] = \frac{\Pr[Data \mid Tree] \; x \; \Pr[Tree]}{\Pr[Data]}$$

Pr [Tree | Data]: posterior probability of the tree.

Pr [Data | Tree]: Conditional likelihood.

Pr [Tree]: Prior probability of the tree.

Pr [Data]: Total probability.

The posterior probability of the tree is not feasible to be calculated analytically so software like MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) use Markov Chain Monte Carlo (MCMC) to perform simulations of the analysis.

Maximum likelihood methods are considered the most reliable due to be based on statistical support. Even though their main drawbacks are reduced variance and being computational highly demanding, becoming in a time consuming process when there are lot of taxa to analyse.

1.6.2 Non parametric methods

Maximum Parsimony is considered as a non parametric method, although sometimes it can input a matrix. Maximum Parsimony finds the topology which requires fewer substitutions among the different sites, the minimum evolutionary tree. This principle is applicable in phylogenetics because in evolution changes are quite rare in a relative short time frame (Xiong, 2006). The algorithm first searches for informative sites, these are sites where the taxa to study exhibit more than two different amino acids and generates the minimum evolution tree in each. The non informative sites, sites that can be explained for a great number of tree topologies, are discarded from that process. From the different topologies identified in each informative site, the overall minimum evolutionary tree is extracted. In cases where more than one minimum evolutionary topology is feasible a consensus is performed (Mount, 2008). Maximum Parsimony

performs correctly in general, although it has been reported that heterogeneous sequences can diminish the accuracy of its results (Kolaczkowski and Thomton, 2004) and produce long branch attraction effect, the long branches are clustered together as a result of the random similarities, not due to common ancestry. The most extended softwares performing an heuristic maximum parsimony method are PAUP (Wilgenbusch and Swofford, 2003) and PHYLIP, Protpars application (Felsenstein, 2005).

1.7 Tree Evaluation

Once a tree has been generated it is necessary to statistically evaluate its reliability (Xiong, 2006). The most extended method is bootstrapping, a technique invented by Efron (Efron, 1979) and introduced into phylogenetics by Felsenstein (Felsenstein, 1985). The fundamental of bootstrapping are based on generating thousands of replicates or pseudoreplicates from the original data and evaluate the branches that are preserved in most of the trees infered on those replicates. The non parametric method deletes some columns from the original multiple sequence alignment dataset by duplicating others. Therefore, distinct datasets are generated. Alternatively, the parametric bootstrapping is based on generating replicates based on a model of sequence evolution (Baxevanis and Ouellette, 2004). The different replicates generated have to be examined in order to obtain a consensus tree which would contain the taxa relationships that have been produced in the overall of the replicates. The bootstrap value in the branches exhibit the percentage of appearance of the clade from it derived (Xiong, 2006). Bootstrap can be applied to distance methods, Maximum Parsimony and Maximum Likelihood. However in Bayesian analysis bootstrapping is not required due to the resampling of trees thousand of times, during the procession of the algorithm. Bayesian analysis “credibility values” that are the statistical support calculated in the Bayes methodology. Credibility values tend to be higher than the bootstrap values and are not directly comparable due to the different statistical methods that they are

generated from. It has been reported that a bootstrap value about 70% gives similar accuracy as a credibility value of 95% (Xiong, 2006).

1.8 Phylogenetic Trees

Phylogenetic trees are graphs that depict the evolutionary path of the taxons of study. They are the result provided by the different tree-building methods. Phylogenetic trees are composed by branches and nodes (Figure 2). Branches are considered the edges from one node to another. Depending on whether the branches depict evolutionary distance, the trees are subsequently called dendograms or cladograms. Nodes can be internal or external. Internal nodes represent the Last Common Ancestor (LCA) of the taxa from which they derive, (Baldauf, 2003). Terminal nodes represent the Operational Taxonomics Units (OTUs), the different sequences analysed. In order to describe a tree the user might compare different groups of OTUs:

- **Monophyletic group:** a group of OTUs descending from a single common ancestor. This is commonly known as a clade.
- **Paraphyletic group:** clade where some descendents have been excluded
- **Polyphyletic group:** association of distantly related OTUs

The structure of the phylogenetic tree is defined as the tree topology. Tree topology is variable as the same tree, derived from the same dataset and method, can present different structures. That is due to the fact that the branches can rotate around the nodes (Baldauf, 2003). The topology of one tree can be rooted or unrooted. The root is considered the ancestral point of the tree and is usually an outgroup differentiated from the ingroup of study. For example when one is comparing mammalian orthologues, a non-mammalian orthologue may be included in the analysis as the outgroup. Unrooted trees lack the defined “rooted” node. Therefore the unrooted trees depict the evolutionary relationships between the taxons of study, but there is no common ancestor to

the whole group. In order to transform one unrooted tree to rooted one it in necessary to define the outgroup (Whelan, 2008).

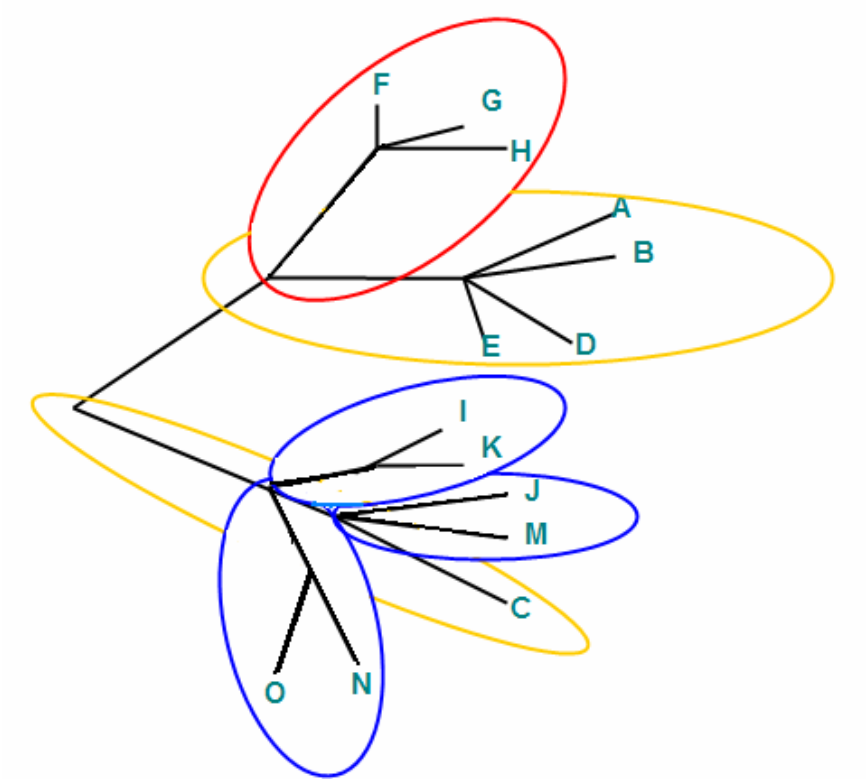


Figure 2. Phylogenetic tree schema.

***Red:** Monophyletic group containing F, G and H taxa. **Blue:** Paraphyletic group, containing I, K, J, M, O and N taxa. **Yellow:** Poliphyletic group: containing A, B, D, E and C taxa.*

1.9 Phylogenetics in drug discovery

Molecular phylogenetics is frequently used at the start of the drug discovery process (Figure 3). Phylogenetic analysis of different related genes may infer the protein function. This technique, complemented with other bioinformatics methods, serves to narrow the initial amount of drug candidates in a pharmacological experiment. In addition, molecular phylogenetics is highly relevant for the process from choosing an appropriate gene to target to elucidating relevant animal models for drug testing. Before a drug is tested in clinical human trials, it is necessary to have been tested in animals and have reported favourable results (GlasxoSmithKline Pharmaceuticals, 2008). However a drug may have different responses when acting in different species due to changes in the gene sequence, potentially resulting in an altered protein structure or function. Robust phylogenetic analysis for a drug target in diverse animal models can identify those most similar to human and therefore, the most appropriate animal model.

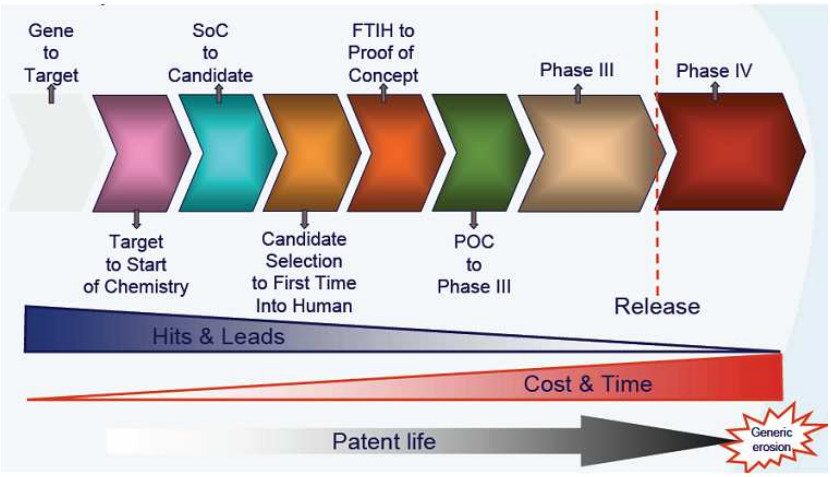


Figure 3. Drug discovery process diagram.

Figure courtesy of Matt Hall, GlaxoSmithKline Pharmaceuticals R & D.

1.10 Aims and objectives

The overall aims of this thesis were to emphasise the relevance of phylogenetic analysis in drug discovery, the necessity of performing more than one phylogenetic method in the analyses -robust analyses- and to improve the use of robust analyses in GlaxoSmithKline Pharmaceuticals (GSK).

In order to address these aims the objectives of the thesis here presented were to introduce a brief explanation of phylogenetic analysis, the most common tree building methods used and its relevance (Chapter 1). Secondly, to perform a robust phylogenetic analysis using four distinct tree building methods, Neighbor Joining, Maximum Parsimony, Maximum Likelihood and Bayesian, in a protein family of great interest in drug discovery -the ABC transporters superfamily- in different species used as pharmacological trials - (mouse, rat, dog and chicken)- (Chapter 2). And finally, to generate a programme to link up the different statistical packages used in Chapter 2 so that the Computational Biology department of GSK could run robust phylogenetic analysis automatically (Chapter 3).

Chapter 2. Phylogenetic analysis of the ABC transporter superfamily

This chapter describes the phylogenetic analysis manually performed on the ATP-Binding Cassete transporters superfamily- potential drug targets- in different species of drug discovery process relevance - human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), dog (*Canis familiaris*) and Chicken (*Gallus gallus*).

2.1 Introduction

The ATP-Binding Cassette (ABC) transporters constitute the largest superfamily of membrane transporters and in fact one of the largest of all protein families (Dean *et al.*, 2001b). The principal function of the ABC transporter superfamily members is to unidirectionally translocate substrates, such as cell nutrients and toxins, across the cell membrane (Dean *et al.*, 2001a; Davidson and Maloney, 2007). Unlike other transporter protein families, ABC's can be found on organelle or cell membranes (Dean *et al.*, 2001b).

To carry out translocation, ABC transporters use the energy released by an ATP hydrolysis reaction that produces a structural change in the protein to allow the passage of the substrate across the membrane (Higgins, 2001). All eukaryotic ABC transporters possess a basic domain structure that comprises nucleotide-binding fold (NBF) and transmembrane domains (TM) (Figure 4). The NBF are highly conserved domains that bind to ATP and captures the energy released by its hydrolysis (Higgins, 2001; Heimer *et al.*, 2002; Igarashi *et al.*, 2004). The NBF is composed of 3 distinct motifs: Walker A - a phosphate-binding loop; Walker B - a magnesium-binding site, and the signature LSGQ motif, which is unique to ABC transporters (Dean *et al.*, 2001b; Higgins, 2001). The role of the TM domain is to recognise the substrate, thus encoding

substrate specificity, and undertake its translocation from across the membrane. The TM domains are composed of alpha-helices that span the membrane between 6 to 11 times (Davidson and Maloney, 2007).

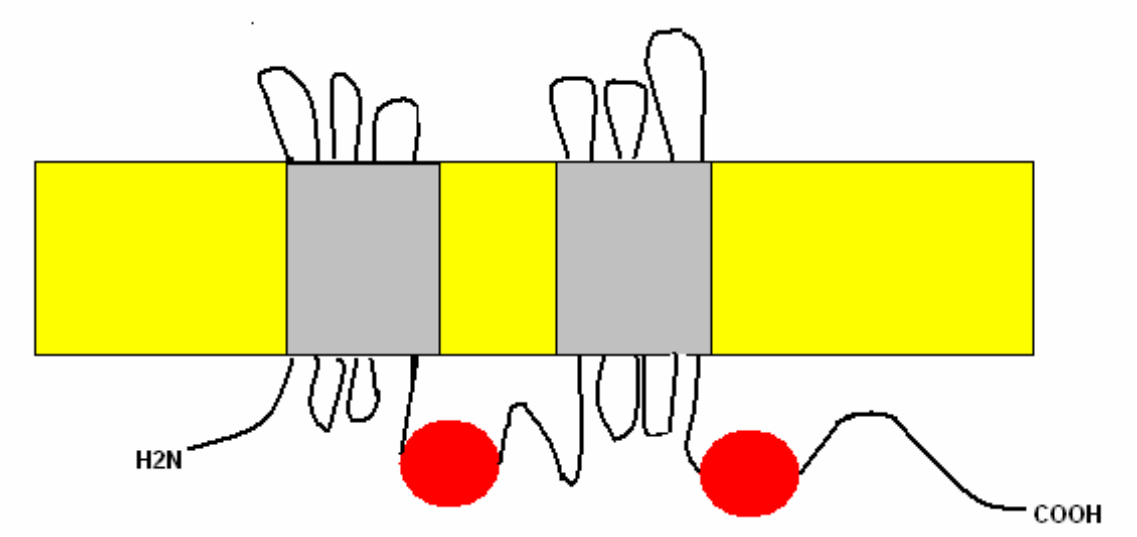


Figure 4. Stylised view of a generic ABC transporter

The coloured text represents the following: Yellow: Lipid bilayer, Grey: Transmembrane domains (TM) and Red: nucleotide-binding folders (NBFs).

Classification of the 48 ABC human transporters has been based on the conservation of the NBF sequences (Dean *et al.*, 2001a), which have split the ABC eukaryotic proteins into seven subfamilies: ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, and White (Dean and Annilo, 2005). These subfamilies have also the name from ABCA to ABCG respectively. The classification of the ABC transporter superfamily also relates to the overall structure of the protein being either “full-length” or “half-length” (Table 2). The full-length protein subfamilies contain two NBFs and two TMs, as presented in Figure 4. The half-length protein subfamilies contain single NBF and TM domains and do not act as active transporters unless they dimerise with a second half-length transporter protein (Higgins, 2001). The half-length transporters can be classed as homodimers or heterodimers depending on whether they bind to members of the same or different subfamily.

Table 2. Classification of the ABC transporters subfamilies

Subfamily	Number of transporters	Structure
ABCA / ABC1	11	Full-length
ABCB / MDR / TAP	7 4	Half-length Full-length
ABCC / MRP	13	Full-length
ABCD / ALD	4	Half-length
ABCE / OABP	1	Full-length
ABCF / GCN20	3	Full-length
ABCG / White	6	Half-length

Members of the ABC transporter superfamily have been associated with diseases such as Alzheimer's disease (Pahnke *et al.*, 2008), cancer (Jamroziak and Robak, 2008; Mizutani *et al.*, 2008; Szakács *et al.*, 2008) and diabetes (Koehn *et al.*, 2008), as well as fifteen severe genetic disorders (Mourez *et al.*, 2000), such as Dubin-Johnson syndrome, Tangier's disease and Stargardt disease (for more details see Appendix II). Furthermore, some ABC transporter family members act as drug efflux pumps (MRP subfamily) (Toyoda *et al.*, 2008). This may also be a common feature of many of the ABC transporter family members whose function has not currently been elucidated. Hence the ABC transporter superfamily is of great relevance for drug discovery.

The ABC transporter superfamily is represented in the majority classes of living organisms (Davidson and Maloney, 2007; Annilo *et al.*, 2006), and contains the largest family of paralogues currently identified (Bouige, 2008). Several phylogenetic analyses of the ABC transporter superfamily have been published and have demonstrated that in the prokaryotes, family members separate by polarity (Saurin *et al.*, 1999) and substrate specificity (Tam and Saier Jr., 1993).

However, a detailed understanding of the relationships within the vertebrate ABC transporters has previously not been resolved. Further, phylogenetic analysis of the vertebrate ABC transporters is highly relevant for the drug discovery process in order to identify appropriate animal models and to help in the identification of new drug targets.

2.2 Aims and Objectives

The overall aim of this chapter is to generate a phylogenetic analysis of the ABC transporter superfamily for the species of interest to drug discovery: *Homo sapiens, Rattus norvegicus, Mus musculus, Canis familiaris and Gallus gallus*.

The objectives of this chapter were to identify the human orthologue gene sequences of the species of interest, to manually run the phylogenetic methods of Neighbor Joining, Maximum Parsimony, Maximum Likelihood and Bayesian with different software packages and to identify and understand the different parameters required in each application. Finally, to generate a consensus tree from the results obtained from each methodology and extract and synthesise the information provided by the results compared.

2.3 Methods

2.3.1 Orthologue identification

The human ABC transporter protein sequences were used to interrogate the nucleotide collection databases of *Pan troglodytes*, *Ratus norvegicus*, *Mus musculus*, *Canis familiaris* and *Gallus gallus* using TBLASTN hosted at the NCBI (Altschul *et al.*, 1997). The default parameters were used for the TBLASTN analyses with an alteration to the expectancy threshold value which was changed into 1 (Altschul *et al.*, 1997; Huxley-Jones *et al.*, 2005) (Table 3). Hits were checked using reciprocal BLAST and by identifying the NCBI Gene name and Gene id. The protein sequence of the orthologues were retrieved and put into a FASTA formatted file.

Table 3. BLAST algorithm parameters used

BLAST ALGORITHM PARAMETERS	
GENERAL	
Max Target sequences	100
Expect Threshold	1
Word size	3
SCORING PARAMETERS	
Matrix	Blosum62
Gap Costs Existence	11 Extension :1
Compositional adjustments	Conditional compositional score matrix adjustment
FILTERS AND MASKING	
Filter: Low complexity regions	No selected
Mask: Mask for look up table only	No selected
Mask: Mask lower case letters	No selected

2.3.2 Alignment

The ABC transporter genes were aligned using CLUSTALX (Thompson *et al.*, 1997) with a BLOSUM62 weight matrix (Henikoff and Henikoff, 1992). The alignment was subsequently stripped for gaps using GeneDOC (Nicholas *et al.*,2007).

2.3.3 Phylogenetic Analysis

Four independent methods were used for phylogenetic analysis: Neighbor Joining, Maximum Parsimony, Maximum Likelihood and Bayesian analysis. Neighbor Joining phylogenetic trees were inferred using CLUSTALX (Thompson *et al.*, 1997) using the Saitou and Nei algorithm (Annilo *et al.*, 2003) and default parameters. Maximum Parsimony phylogenetic trees were inferred using the Seqboot, Protpars and Consense applications in PHYLIP (Felsenstein, 2005). The Seqboot application generates 1000 replicates of the original input dataset, for each of which Protpars infers a phylogenetic tree. The input order of the sequences into Protpars was randomised using a seed of 5 and jumbled 3 times. A consensus bootstrapped tree was generated from the 1000 replicates using the Consense application. Other parameters used in the analyses were the default. Maximum Likelihood phylogenetic trees were inferred using TreePuzzle (Schmidt *et al.*, 2002) and PHYLIP (Felsenstein, 2005). TreePuzzle was used to generate an alpha value from the initial alignment using 8 gamma rate categories. Other parameters used in TreePuzzle were default. The alpha value was subsequently used by the PHYLIP ProML application (Felsenstein, 2005) in order to incorporate heterogenic rates of evolution across residues in the alignment into the Maximum Likelihood algorithm. The alpha value was converted into the gamma value (coefficient of variation) according to the following equation:

$$CV = 1 / a^{1/2}$$

The number of categories used for the in Hidden Markov Models in PROML was 6 (Felsenstein, 2005). Bayesian phylogenetic trees were inferred using MrBayes version 3.0 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Mr Bayes uses a Markov Chain Monte Carlo algorithm as an analytical technique to estimate an approximation of the posterior reliability of the trees. The parameters chosen were 100,000 generations to the algorithm on six independent runs (chains). 50% of these trees were sampled for subsequent consensus analysis. A consensus tree with clade credibility values

was generated using the sumt command with exclusion (burn-in) of the first 1000 trees. Phylogenetic trees obtained from the different tree building methods were visualised and edited in NJplot (Perrière and Gouy, 1996).

2.4 Results

2.4.1 ABC transporter identification

Members of the ABC transporter superfamily in *Homo sapiens, Pan troglodytes, Rattus norvegicus, Mus musculus, Canis familiaris and Gallus gallus* were identified using BLAST analyses. The total amount of sequences retrieved for each animal and subfamily are presented in Table 4. A more detailed relation of the ABC subfamily orthologues in *Homo sapiens, Pan troglodytes*, *Rattus norvegicus*, *Mus musculus*, *Canis familiaris* and *Gallus gallus* is shown in the Tables from 4 to 8. Information relating to the sequence origins and accession numbers is presented in Appendix I.

Previous studies on the evolution of the ABC transporter family (Igarashi *et al.*, 2004; Annilo *et al.*, 2006) have stated that the ABC transporters are represented in most of the phyla of eukaryotes and prokaryotes. The orthologue retrieval analysis agrees with previous studies as all the ABC subfamilies are represented in the six species of interest (Table 4). Analysis of the ABC transporter family as a whole reveals that there are many more ABC-A, B and C-type transporters and that the numbers within the subfamilies are lineage specific (Table 4). ABCE is the only subfamily that maintains the same number of genes in the selected species. ABCD and ABCG maintain the same number of representants in mammals, and the subfamilies ABCB and ABCG depict to have an extra transporter in rodents.

Table 4. The ABC transporter family

The numbers of sequences identified in *Homo sapiens, Pan troglodytes, Rattus norvegicus, Mus musculus, Canis familiaris* and *Gallus gallus* for the different ABC transporters subfamilies are presented.

Subfamily	*Homo sapiens*	*Pan troglodytes*	*Rattus norvegicus*	*Mus musculus*	*Canis familiaris*	*Gallus gallus*
ABCA	14	13	15	14	11	9
ABCB	11	10	12	12	9	6
ABCC	13	12	11	11	11	11
ABCD	4	4	4	4	4	3
ABCE	1	1	1	1	1	1
ABCF	3	3	3	3	3	2
ABCG	5	5	6	6	5	3
Total	**51**	**48**	**52**	**51**	**44**	**35**

The ABCA subfamily

No species presented in Table 5 exhibits the complete complement of ABCA transporters, with chicken (*Gallus gallus*) clearly exhibiting the fewest ABCA orthologues. The variability in gene number within the species is clear when one considers that there are two copies of ABCA8 in rodents; ABCA10 is only present in primates; ABCA14 and ABCA15 have been only identified in rodents; and ABCA11 and ABCA17 are pseudogenes (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008; Piehler *et al.*, 2006).

Table 5. Orthologue representation in the ABCA subfamily

Family ABCA						
ABC transporter	***Homo sapiens***	***Pan troglodytes***	***Rattus norvegicus***	***Mus musculus***	***Canis familiaris***	***Gallus gallus***
ABCA1	x	x	x	x	x	x
ABCA2	x	x	x	x	x	x
ABCA3	x	x	x	x	x	x
ABCA4	x	x	x	x	x	x
ABCA5	x	x	x	x	x	x
ABCA6	x	x	x	x	x	
ABCA7	x	x	x	x	x	
ABCA8	x	x	(a/b)	(a/b)	x	x
ABCA9	x	x	x	x	x	
ABCA10	x	x				
ABCA11	x		x	x		x
ABCA12	x	x	x	x	x	x
ABCA13	x	x	x	x	x	x
ABCA14			x	x		
ABCA15			x	x		
ABCA17	x	x	x			

The ABCB subfamily

In comparison to the ABCA family there is consistency in the number of ABCB genes in the studied species (Table 6). All the ABCB transporters present in human have an orthologue in rat and mouse. However it is of note that there are two ABCB1 orthologues in rodents (a/b) (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008; Kalabis *et al.*, 2005). The main differences in ABCB complement within the mammalian lineage appear to be the absence of ABCB4 and ABCB10 in dog and ABCB7 in chimpanzee.

Table 6. Orthologue representation in the ABCB subfamily

Family ABCB						
ABC transporter	***Homo sapiens***	***Pan troglodytes***	***Rattus norvegicus***	***Mus musculus***	***Canis familiaris***	***Gallus gallus***
ABCB1	x	x	(a/b)	(a/b)	x	x
TAP1 (ABCB2)	x	x	x	x	x	
TAP2 (ABCB3)	x	x	x	x	x	
ABCB4	x	x	x	x		x
ABCB5	x	x	x	x	x	
ABCB6	x	x	x	x	x	x
ABCB7	x		x	x	x	x
ABCB8	x	x	x	x	x	
ABCB9	x	x	x	x	x	x
ABCB10	x	x	x	x		x
ABCB11	x	x	x	x	x	

The ABCC subfamily

The family ABCC is highly conserved across the species presented in Table 7, with the ABCC11 and ABCC13 genes as exceptions. The lack of ABCC13 in mammals apart from human has previously been elucidated as to pseudogenisation in mammals (Annilo and Dean, 2004).

Table 7. Orthologue representation in the ABCC subfamily

Family ABCC						
ABC transporter	***Homo sapiens***	***Pan troglodytes***	***Rattus norvegicus***	***Mus musculus***	***Canis familiaris***	***Gallus gallus***
ABCC1	x	x	x	x	x	x
ABCC2	x	x	x	x	x	x
ABCC3	x	x	x	x	x	x
ABCC4	x	x	x	x	x	x
ABCC5	x	x	x	x	x	x
ABCC6	x	x	x	x	x	x
CFTR (ABCC7)	x	x	x	x	x	x
ABCC8	x	x	x	x	x	x
ABCC9	x	x	x	x	x	x
ABCC10	x	x	x	x	x	x
ABCC11	x	x			x	
ABCC12	x	x	x	x	x	
ABCC13	x					x

The ABCD, ABCE, ABCF and ABCG subfamilies

There is high conservation of the complement of ABCD, ABCE, ABCF and ABCG transporters across the six species; however it is apparent that ABCG3 is only present in rodents (Table 8).

Table 8. Orthologue representation in the ABCD, E, F and G subfamilies

ABC transporter	*Homo sapiens*	*Pan troglodytes*	*Rattus norvegicus*	*Mus musculus*	*Canis familiaris*	*Gallus gallus*
Family ABCD						
ABCD1	x	x	x	x	x	
ABCD2	x	x	x	x	x	x
ABCD3	x	x	x	x	x	x
ABCD4	x	x	x	x	x	x
Family ABCE						
ABCE1	x	x	x	x	x	x
Family ABCF						
ABCF1	x	x	x	x	x	
ABCF2	x	x	x	x	x	x
ABCF3	x	x	x	x	x	x
Family ABCG						
ABCG1	x	x	x	x	x	x
ABCG2	x	x	x	x	x	x
ABCG3			x	x		
ABCG4	x	x	x	x	x	
ABCG5	x	x	x	x	x	x
ABCG8	x	x	x	x	x	

2.4.2 Phylogenetic Analysis

From the initial non gap-stripped alignment the chimpanzee (*Pan troglodytes)* orthologues were discarded as they introduced too many gaps in the alignment. A robust gap-stripped alignment of the ABC transporter family in human, dog, chicken, mouse and rat (*Homo sapiens, Canis familiaris*, *Gallus gallus*, *Mus musculus* and *Rattus norvegicus*) could not be generated as the phylogenetic signal was lost from gap-stripping.

It is clear from the phylogenetic analysis of the whole ABC transporter family (Figure 5) that the ABCA genes are more divergent to the rest of ABC transporters. There is statistically supported evidence for the divergence of the F and GE clades, from the BCD clade. However, due to the lack of statistical support for the branchpoint topology between the B, C and D clades, these clades have been collapsed. It is not possible to determine the evolutionary relationships among the ABCB, ABCC and ABCD subfamilies due to their high sequence similarity.

In order to generate further insight into the true topology among the collapsed clades a bootstrapped Neighbour Joining phylogenetic tree was inferred from an alignment of the human, mouse and rat ABC transporter sequences (Digital Appendix). However, the relationships between the clades were again not robust (bootstrap values inferior to 70%) and as a result no topology refinement among the collapsed groups could be performed.

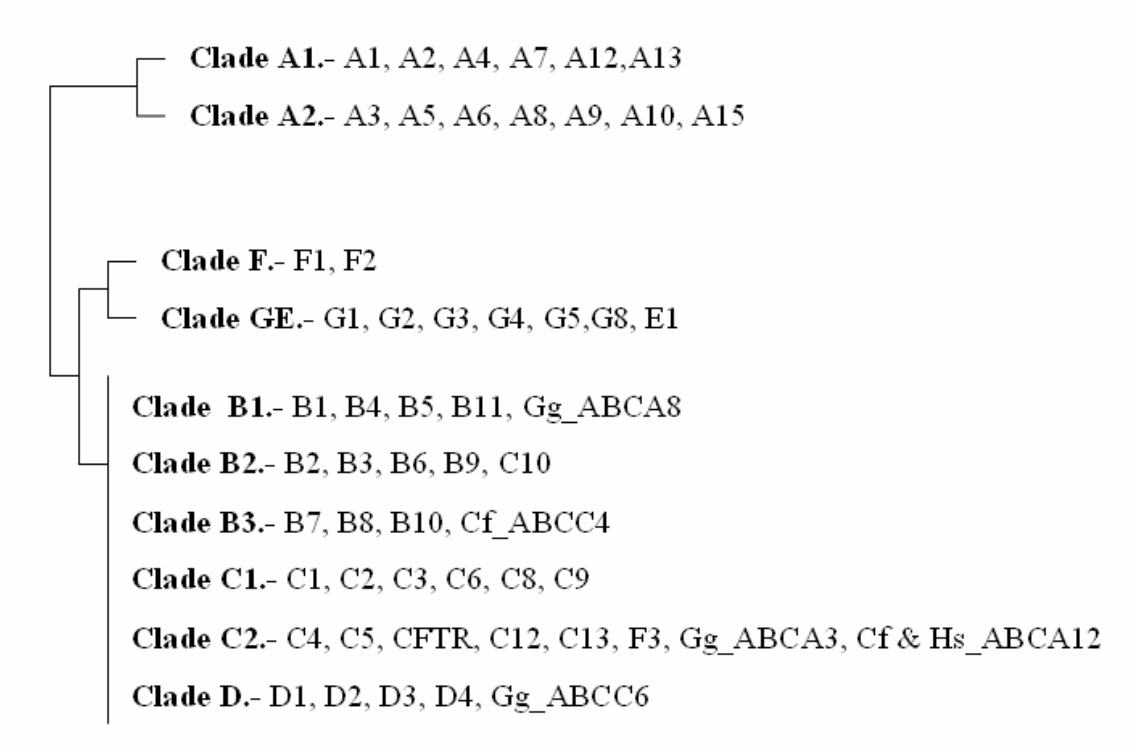


Figure 5. Phylogenetic relationships of the ABC transporters family.

The family was separated into 10 sub-groups based on a strong bootstrap from a Neighbor Joining gapped tree of the ABC transporters in human, chicken, dog, mouse and rat. The relationships between clades were extracted from a Neighbor Joining gapped tree where only human ABC transporters were used Due to higher statistical support in their branching.

Therefore, based on the initial non gap-stripped Neighbor Joining tree of the ABC transporters in human, rat, mouse, dog and chicken (Figure 5), the ABC transporter family was split into 10 groups, each supported by significant bootstrap values. For each group a gap-stripped alignment was generated and multiple phylogenetic analyses were inferred (Figure 6 to Figure 15). As a general overview, it can be said that all the clades preserve taxonomic topology – for example mouse and rat sequences are clustered together. Chicken (*Gallus gallus*) was used as an outgroup to the mammalian sequences. All phylogenetic trees generated are presented in Digital Appendix.

CLADE A1

Phylogenetic analysis was performed on the ABC A1 clade, containing the ABCA1, ABCA4, ABCA7, ABCA12 and ABCA13 genes (Figure 6). There is strong statistical support between the orthologues, however there is little or no support between paralogues. It is striking that the ABCA7 orthologues are not clustered together, with the human and mouse sequences being placed at the root of the tree and the rat sequence being located between the ABCA2 and the ABCA1/4 clades. Although only one method, Neighbor Joining, produced statistical support for the position of the rat sequence, the other methods (Maximum Parsimony, Maximum Likelihood and Bayesian analysis) also place the rat sequence away from its orthologues (Digital Appendix).

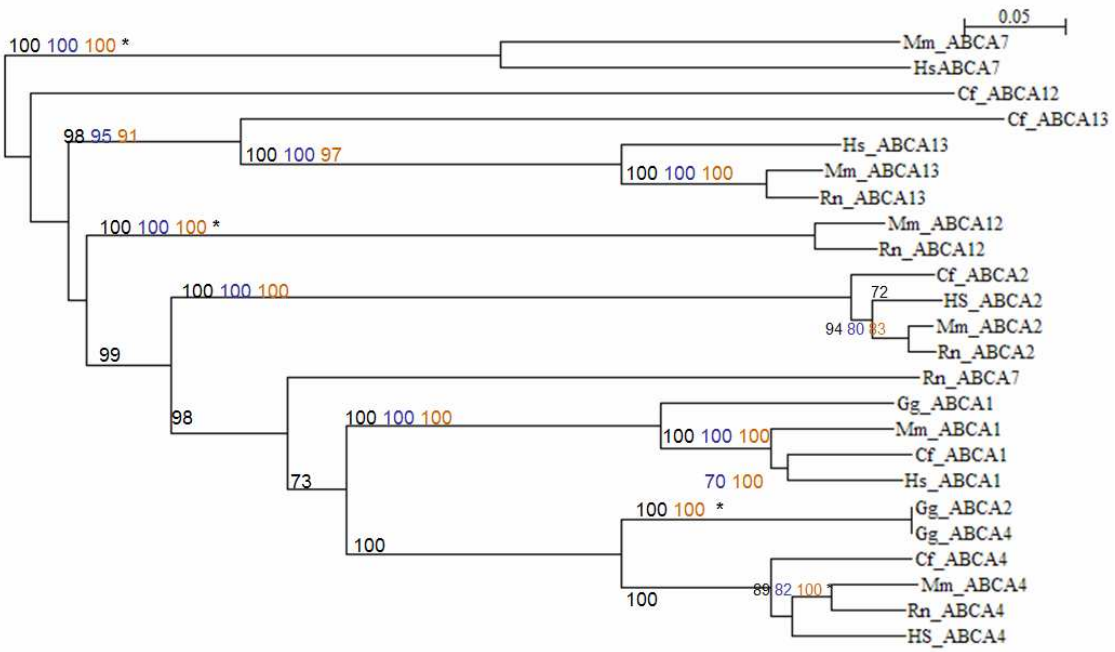


Figure 6. Phylogenetic analysis of the A1 clade.

The tree presented was inferred using Neighbor Joining. Confidence for each branchpoint is presented by bootstrap values from Neighbor Joining (black) and Maximum Parsimony (blue). Bayesian clade credibility values are depicted in brown text. Clades that agree in the Maximum Likelihood phylogeny are indicated by an asterix. Species are annotated by Cf – dog, Gg – chicken Hs – human, Mm – mouse and Rn – rat.

CLADE A2

Phylogenetic analysis performed on the ABC A2 clade containing A3, A5, A6, A8, A10, A15 genes, reveals two distinct clades (Figure 7). The clade containing the ABCA3, ABCA14, ABCA15 and ABCA17 genes is named the A2a clade (Figure 7). Unlike the ABCA3 orhtologues, ABCA14, ABCA15 and ABCA17 are not present in human. It is of note that ABCA17 is a pseudogene in man, but no evidence of ABCA14 and ABCA15 pseudogenes could be detected. Previous phylogenetic analyses have also placed the ABCA14, ABCA15 and ABCA17 genes closer to ABCA3 (Chen *et al.*, 2004). It is of note that the rat ABCA15 shows a high diversification by long branch lengths. The clade containing the ABCA5, ABCA6, ABCA8, ABCA9 and ABCA10 is named the A2b clade and shows well supported branching ABCA5, ABCA6, ABCA8 and ABCA9 all have orthologues present in the model organisms (Figure 7). It is of note that the rodent ABCA8a genes show a high level of sequence diversification away from the other ABCA8 orthologues. In addition, interrogation of non-human primate genomes confirms the fact that ABCA10 is only present in primate genomes (Table 5).

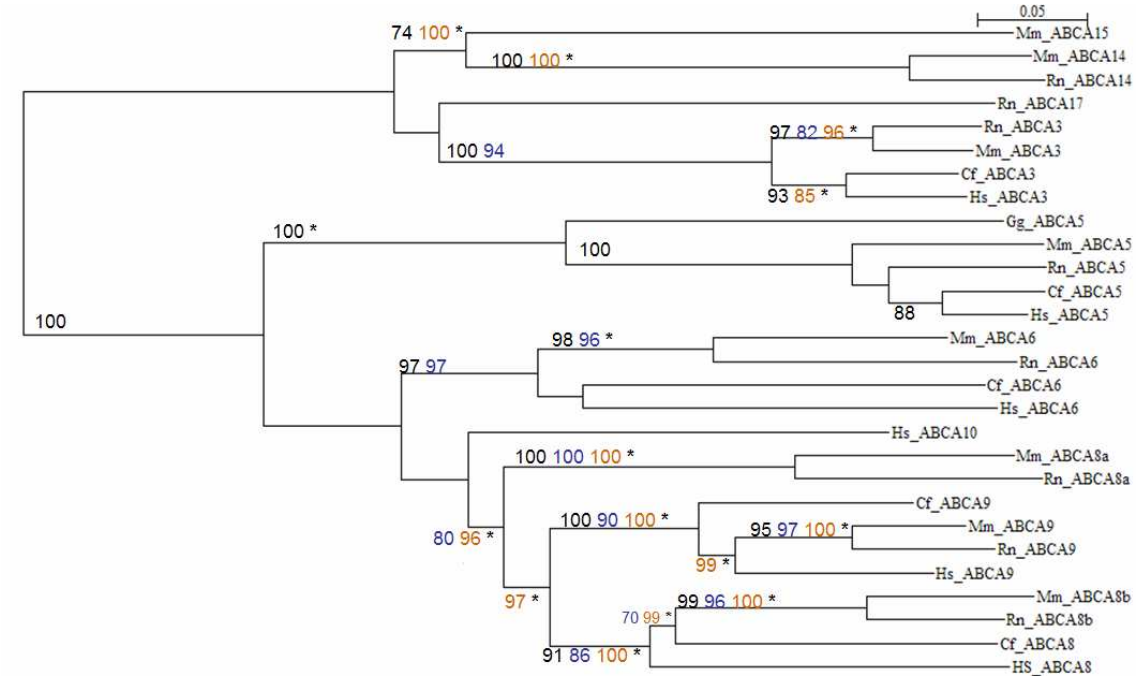


Figure 7. Phylogenetic tree of the A2 clade

The tree presented was inferred using Neighbor Joining. Confidence for each branchpoint is presented by bootstrap values from Neighbor Joining (black) and Maximum Parsimony (blue). Bayesian clade credibility values are depicted in brown text. Clades that agree in the Maximum Likelihood phylogeny are indicated by an asterix. Species are annotated by Cf – dog, Gg – chicken Hs – human, Mm – mouse and Rn – rat.

CLADE F

The ABCF subfamily is based on the genes ABCF1, ABCF2 and ABCF3. However, no statistical support was found in the original Neighbor-Joining tree to group ABCF3 with the rest of the members of the family. Therefore, the clade F depicts then the evolution of F1 and F2 (Figure 8). There is no support for the order of the branching between the orthologues of the ABCF2 genes, probably due to the high levels of sequence similarity between these sequences (Figure 8).

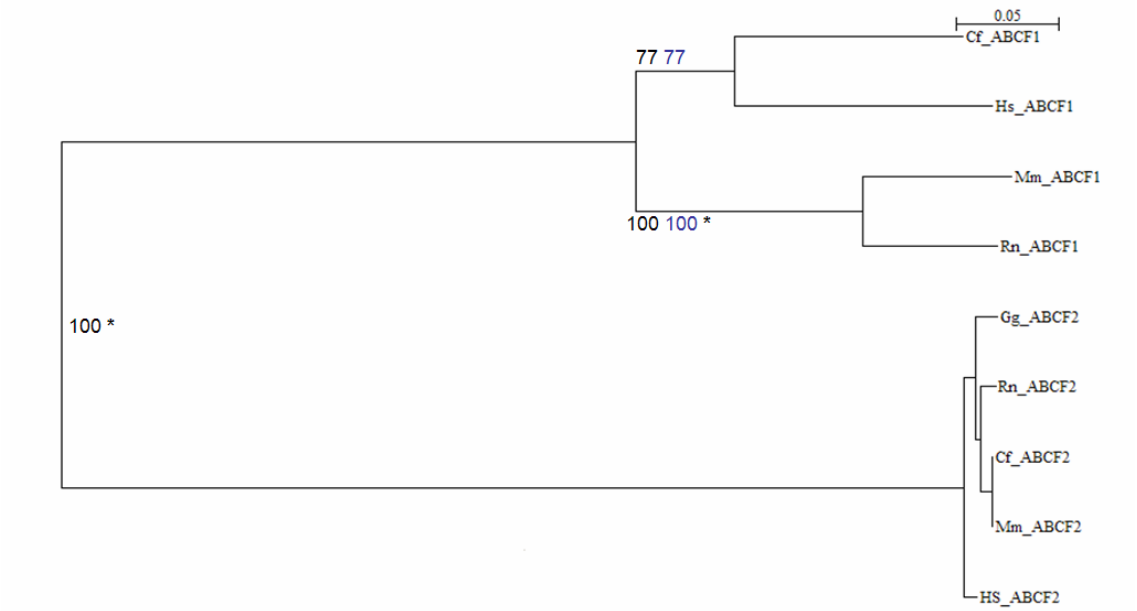


Figure 8. Phylogenetic tree of the F clade

The tree presented was inferred using Neighbor Joining. Confidence for each branchpoint is presented by bootstrap values from Neighbor Joining (black) and Maximum Parsimony (blue). Bayesian clade credibility values are depicted in brown text. Clades that agree in the Maximum Likelihood phylogeny are indicated by an asterix. Species are annotated by Cf – dog, Gg – chicken Hs – human, Mm – mouse and Rn – rat.

CLADE GE

The GE clade contains the ABCG1, ABCG2, ABCG3, ABCG4, ABCG5, ABCG8 and ABCDE genes (Figure 9). There was statistical support in the Neighbor Joining ungapped tree for these sequences to be grouped together into one clade for subsequent phylogenetic analysis. However, it is clear from more refined phylogenetic analysis (Figure 9) that the ABCE genes are outlyers to this family. There are single orthologues in each species for most of the ABCG families apart from the rodent ABCG2/3 clade, where the mouse ABCG2 and G3 sequences group strongly with two rat ABCG3 genes (ABCG3 and ABCG3s).

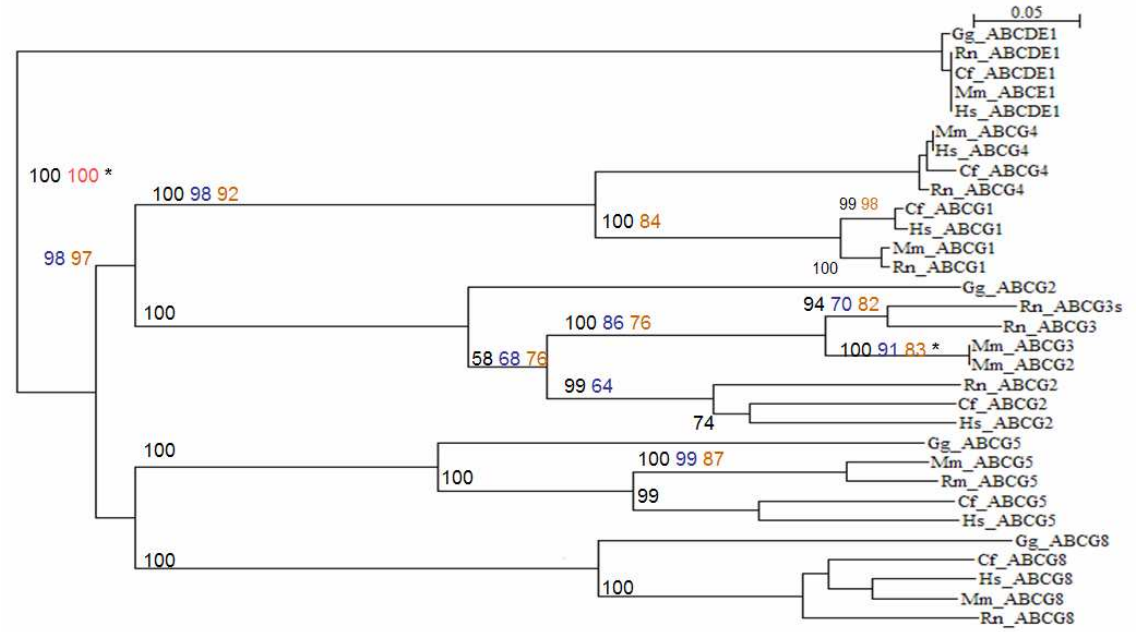


Figure 9. Phylogenetic tree of the GE clade

The tree presented was inferred using Neighbor Joining. Confidence for each branchpoint is presented by bootstrap values from Neighbor Joining (black) and Maximum Parsimony (blue). Bayesian clade credibility values are depicted in brown text. Clades that agree in the Maximum Likelihood phylogeny are indicated by an asterix. Species are annotated by Cf – dog, Gg – chicken Hs – human, Mm – mouse and Rn – rat.

CLADE B1

The ABC B1, B4, B5, B11 and chicken ABCA8 were grouped together from the initial Neighbor-Joining tree (Figure 5) and subsequent phylogenetic analysis was performed (Figure 10). However, there is no statistical support to group the chicken ABCA8 (Gg_ABCA8) sequence with the rest of this clade. In addition, there are duplicates for gene ABCB1 in the rodent lineage.

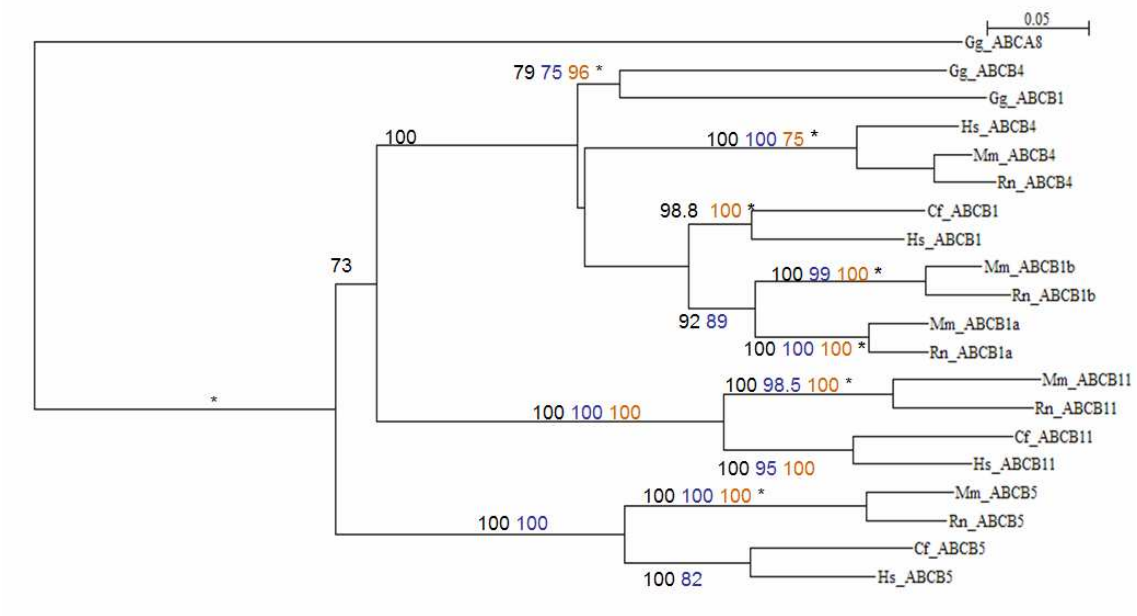


Figure 10. Phylogenetic tree of the B1 clade

The tree presented was inferred using Neighbor Joining. Confidence for each branchpoint is presented by bootstrap values from Neighbor Joining (black) and Maximum Parsimony (blue). Bayesian clade credibility values are depicted in brown text. Clades that agree in the Maximum Likelihood phylogeny are indicated by an asterix. Species are annotated by Cf – dog, Gg – chicken Hs – human, Mm – mouse and Rn – rat.

CLADE B2

Plylogenetic analysis of the B2 clade reveals the evolutionary relationships of ABCB2, ABCB3, ABCB6, ABCB9 and ABCC10 (Figure 11). There is only statistical support from the Neighbor Joining method for the grouping of the ABCB2 and ABCB3 paralogues, and no statistical support is present for the branching order for the other paralogues. It is of note that the dog ABCB6 sequence does not cluster with the other ABCB6 orthologues in any method used (Digital Appendix) and there is no significant bootstrap value to support this grouping.

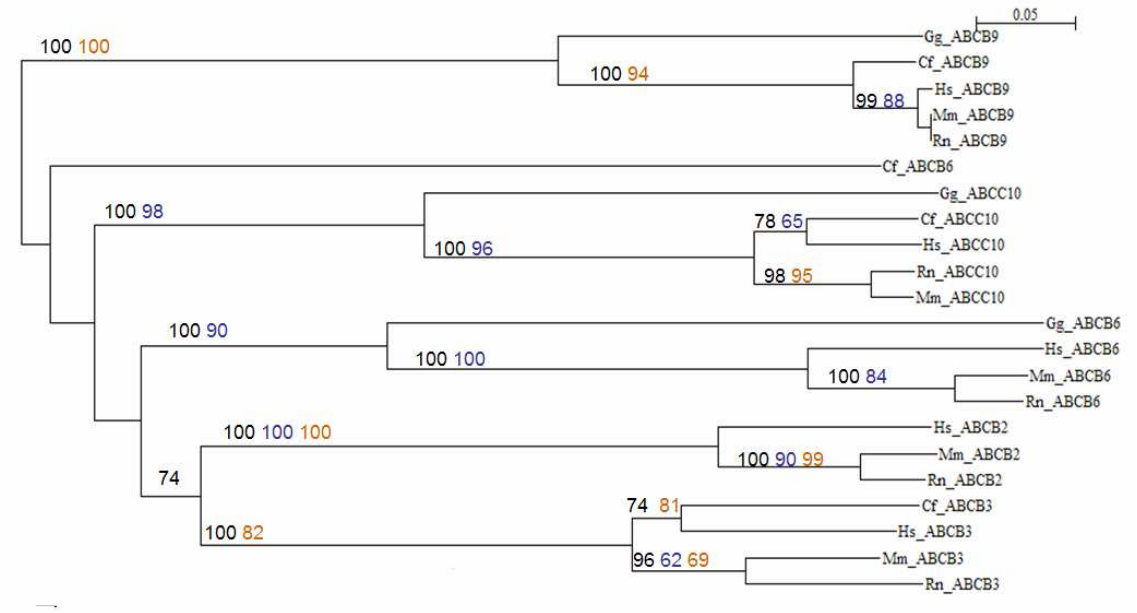


Figure 11 . Phylogenetic tree of the B2 clade

The tree presented was inferred using Neighbor Joining. Confidence for each branchpoint is presented by bootstrap values from Neighbor Joining (black) and Maximum Parsimony (blue). Bayesian clade credibility values are depicted in brown text. Clades that agree in the Maximum Likelihood phylogeny are indicated by an asterix. Species are annotated by Cf – dog, Gg – chicken Hs – human, Mm – mouse and Rn – rat.

CLADE B3

The B3 clade contains the dog ABCC4 and the ABCB7, ABCB8, ABCB10 gene families (Figure 12). Despite being grouped away from the other ABCC4 orthologues (Figure 5) it is striking high support for the grouping of the dog ABCC4 with the ABCB8 genes. However, due to the significantly long branch of the ABCC4 gene, it is likely that the dog ABCC4 gene is still an outlier to this clade.

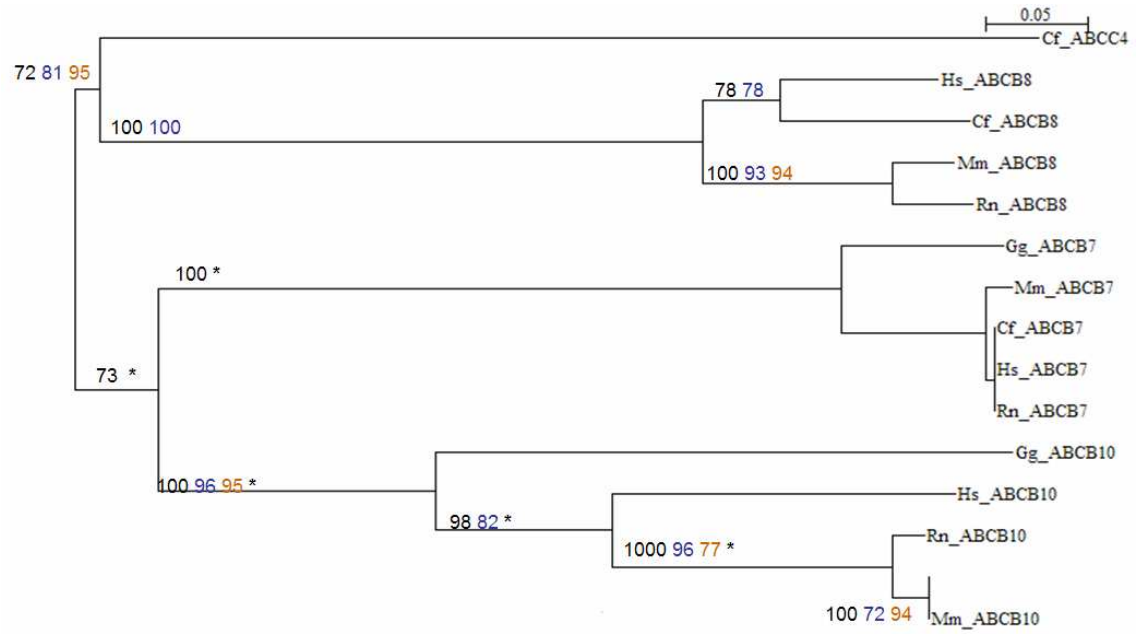


Figure 12. Phylogenetic tree of the B3 clade

The tree presented was inferred using Neighbor Joining. Confidence for each branchpoint is presented by bootstrap values from Neighbor Joining (black) and Maximum Parsimony (blue). Bayesian clade credibility values are depicted in brown text. Clades that agree in the Maximum Likelihood phylogeny are indicated by an asterix. Species are annotated by Cf – dog, Gg – chicken Hs – human, Mm – mouse and Rn – rat.

CLADE C1

The C1 clade contains ABCC1, ABCC2, ABCC3, ABCC6, ABCC8 and ABCC9 (Figure 13). Most of the paralogues contain orthologues from all six species whose branching follows taxonomic topology. However, the human ABCC6 gene is not presented as the sequence is truncated (99 amino acids) further, the chicken ABCC6 gene does not cluster with its mammalian orthologues. In addition it is apparent that ABCC8 and ABCC9 are highly similar.

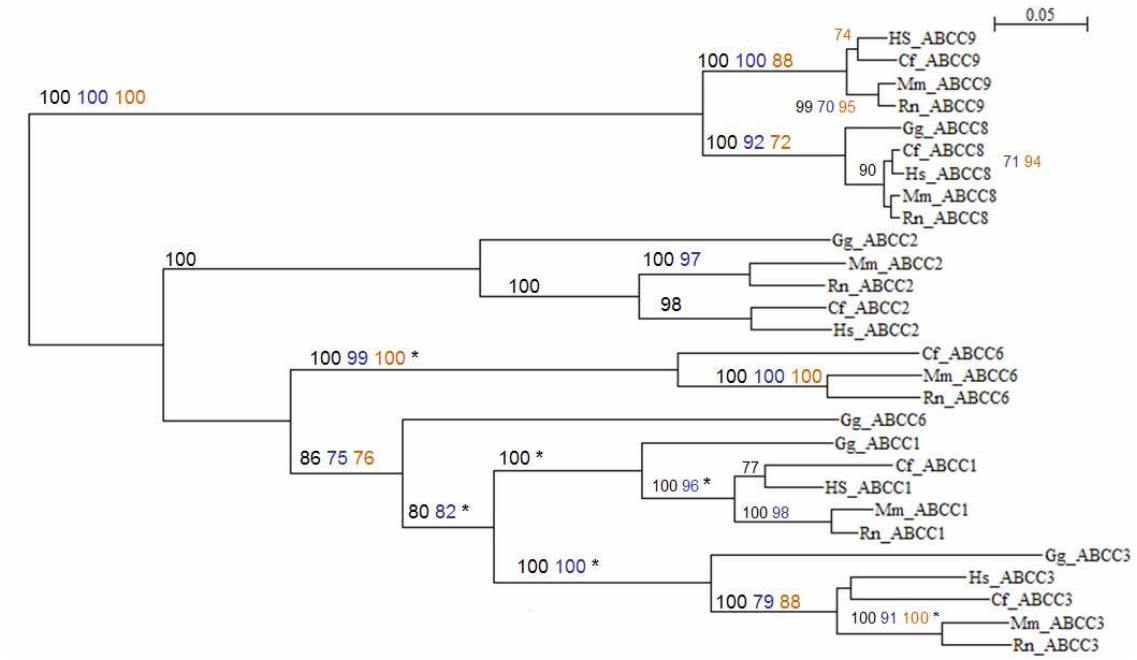


Figure 13. Phylogenetic tree of the C1 clade

The tree presented was inferred using Neighbor Joining. Confidence for each branchpoint is presented by bootstrap values from Neighbor Joining (black) and Maximum Parsimony (blue). Bayesian clade credibility values are depicted in brown text. Clades that agree in the Maximum Likelihood phylogeny are indicated by an asterix. Species are annotated by Cf – dog, Gg – chicken Hs – human, Mm – mouse and Rn – rat.

CLADE C2

The C2 clade contains ABCC4, ABCC5, CFTR, ABCC12, ABCC13, ABCF3, the chicken ABCA3, and the ABCA12 genes (Figure 14). Sequences not belonging to ABCC subfamily show strong statistical support to be clustered together. In addition, the clustering of the ABCC13, ABCC4 and CFTR paralogues is well supported (Figure 14). However, there is no support to maintain the chicken ABCA3 gene (Gg_ABCA3) within C2 clade. Therefore it could have been equally grouped with the rest of ABCA3 sequences.

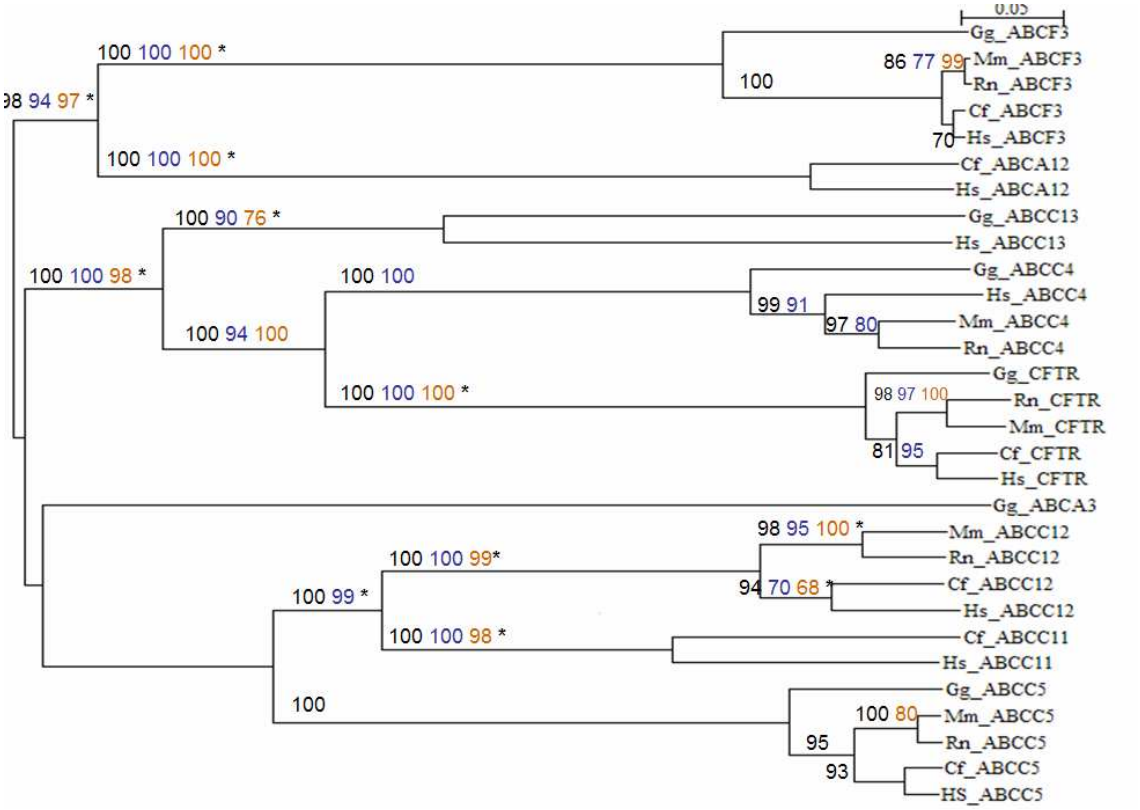


Figure 14. Phylogenetic tree of the C2 clade

The tree presented was inferred using Neighbor Joining. Confidence for each branchpoint is presented by bootstrap values from Neighbor Joining (black) and Maximum Parsimony (blue). Bayesian clade credibility values are depicted in brown text. Clades that agree in the Maximum Likelihood phylogeny are indicated by an asterix. Species are annotated by Cf – dog, Gg – chicken Hs – human, Mm – mouse and Rn – rat.

CLADE D

The D clade contains the ABCD sequences. From the initial Neighbor Joining tree (Figure 5) the chicken ABCC6 sequence weakly clustered with this family, however more refined phylogenetic analysis (Figure 15) reveals that this is an unsupported, and probably false, grouping.

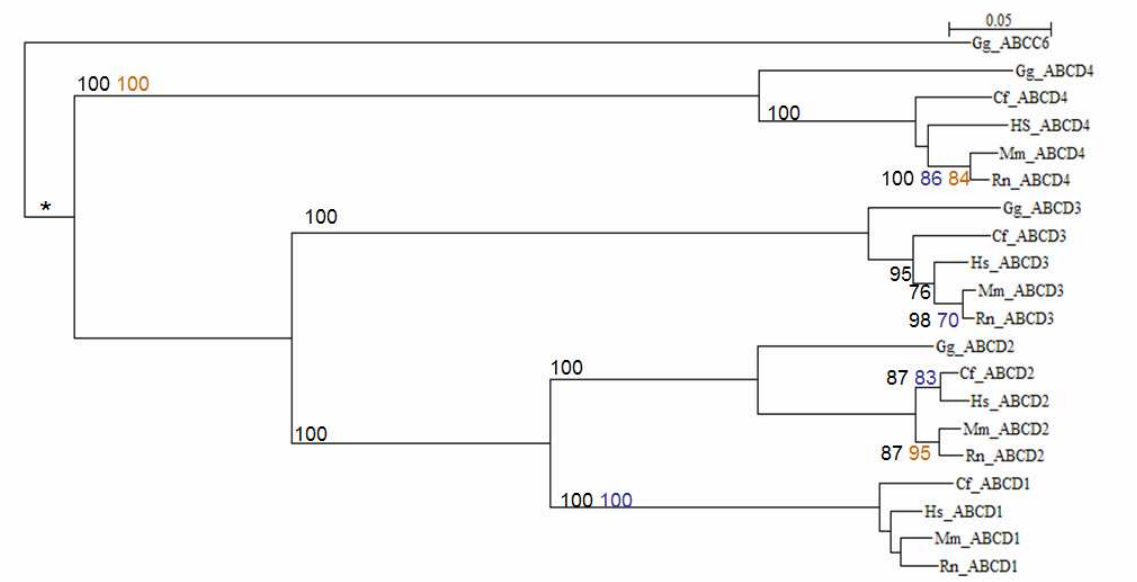


Figure 15. Phylogenetic tree of the D clade

The tree presented was inferred using Neighbor Joining. Confidence for each branchpoint is presented by bootstrap values from Neighbor Joining (black) and Maximum Parsimony (blue). Bayesian clade credibility values are depicted in brown text. Clades that agree in the Maximum Likelihood phylogeny are indicated by an asterix. Species are annotated by Cf – dog, Gg – chicken Hs – human, Mm – mouse and Rn – rat.

2.5 Discussion

The ABC transport superfamily is the largest gene family currently known (Bouige, 2007). Members of all the ABC subfamilies are represented in each of the model organisms studied. It is clear that the number of ABC transporters is lineage specific. Of the species analysed human, mouse and rat genomes have the most ABC genes (Table 4). A possible explanation of why these genomes have more ABC transporters could be due to the fact that they are the most studied organisms. The human genome was sequenced in 2004 and reviewed in 2005, the mouse sequenced in 2002 last reviewed in 2006, and the rat sequenced and updated in 2004 (Ensembl, 2008; Gregory *et al.*, 2002; Rat Genome Sequencing Project, 2004). More recently in 2005 the chimpanzee and the dog genomes were sequenced (The Chimpanzee Sequencing and Analysis Consortium, 2005; Lindblad-Toh *et al.*, 2005), however as genome analysis is an iterative process their sequence state should currently be considered as a draft. This may explain why there is a lack of some transporters such as ABCA11 in chimpanzee and dog, ABCB4 and ABCB10 in dog and ABCB7 in chimpanzee (Table 5 and Table 6). Alternatively the absence of such genes may reflect the plasticity of the ABC complement in mammalian genomes. The chicken genome was started in 2004, reviewed in 2005 and thus may be considered near complete (Altschul *et al.*, 1997; Ensembl, 2008). Therefore, the lack of the ABC transporters in the chicken is less likely to be due to an "incomplete" genome and more likely to be due to its phylogenetic divergence to the rest of the group. Many of these genes may have evolved after the bird phyla diverged from the lineage where mammals descend. Indeed, chicken has the fewest ABC transporters in all the ABC eukaryotic subfamilies (Tables from 4 to 8).

Further, the pseudogenisation of the human ABCA11, ABCA17 and ABCC13 genes (Table 5 and Table 7) represents a dynamic gene-birth process on the ABC transporters. Indeed it is clear that rodents contain the most divergent ABC transporters in the species studied. In most cases the rodent orthologues

group in the correct taxonomic position. However, the positions of ABCA8, ABCA14, ABCA15 and ABCG3 are due to the fact that they are rodent specific. In addition, ABCA8 and ABCB1 are duplicates in rodents (Tables 5 and 6; Figures 7 and 10). Such rodent specific genes also highlight the differences between the mouse and rat genomes versus the human. Such alterations in ABC transporter complement may have implications when using rodents as model organisms as not only do they have more genes, but some direct orthologues are highly divergent in sequence from that of man, thus potentially also exhibiting altered functions.

Abherrant grouping of some genes is present within the phylogenetic analyses presented (Figures 10, 11, 12, 14 and 15). Their misalignment may be due to one or more or the following:

- Misaligned sequences

  The alignment may not align some sequences correctly making them more similar to another family. Some misalignments may have resulted from segments or the entire protein sequence being highly divergent.

- Insertions/ deletions

  Insertions and deletions may carry the phylogentic signal defining where an ABC transporter gene should be place in the tree. Therefore when the gap-stripped alignment was performed, the sequence may have lost its similarity to or “uniqueness” from the rest gene family.

- Fast evolving sequences

  Fast evolving genes accumulate many more mutations than non fast-evolving genes due to either relaxed selection or positive selection pressure on the whole or specific residues of the sequence. In extreme cases such genes may loose some similarity to their orthologues and paralogues thus altering the phylogenetic signal and position in a phylogenetic tree.

- Artifact effect

  Each algorithm for phylogenetic analysis is based on distinct statistics that each may have innate biases. Thus, sometimes errors in branching can be generated. However, performing multiple phylogenies diminishes this problem in some extent.

With regard to such potential misalignments, the tree topology presented in Figure 5 has been subsequently altered and the consensus topology is depicted in Figure 16. This topology (Figure 16), although in agreement with previous simplified phylogenetic analyses (Dean *et al.*, 2001b), defines the ABC superfamily based on robust phylogenetics in multiple species. All the individual clades in this study were analysed using four distinct tree-building methods. The methods selected, Neigbor Joining, Maximum Parsimony, Maximum Likelihood and Bayesian, were chosen because of using the most divergent calculating models of the evolutionary relationships among sequences. Therefore, because no method guarantees finding the correct phylogenetic tree, the more a taxonomic relation is supported by different methods would be considered as a sign of its veracity. The topology of the ABC transporter superfamily demonstrates how the clades are divided by the different subfamilies and that can influence by the different structure presented among them.

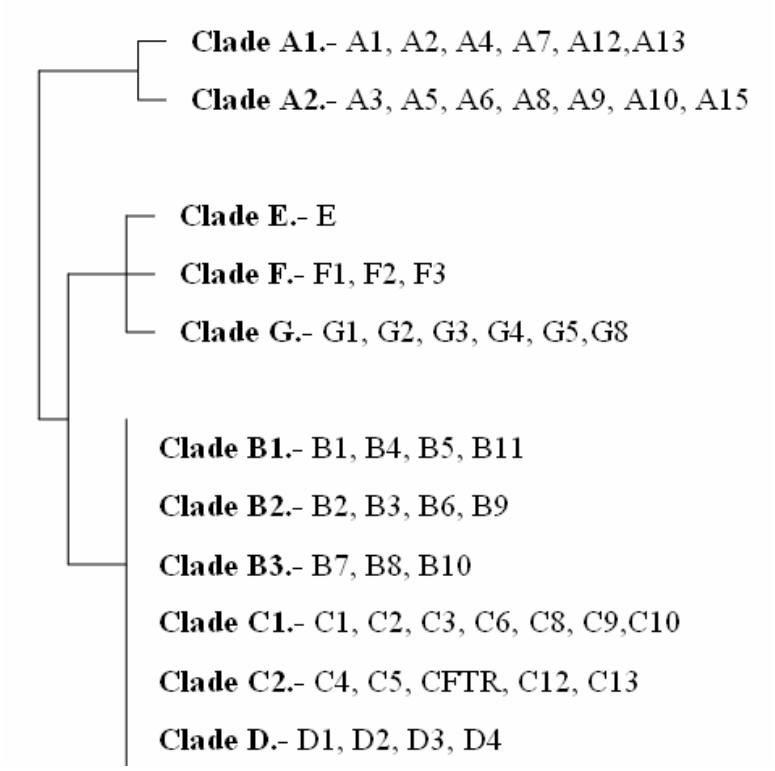


Figure 16. Phylogenetic relationships of the ABC transporter superfamily.

The family was separated into 10 sub-groups based on a strong bootstrap from a Neighbor Joining gapped tree of the ABC transporters in human, chicken, dog, mouse and rat. The relationships between clades were extracted from a Neighbor Joining gapped tree where only human ABC transporters were used Due to higher statistical support in their branching.

ABCA genes

From the A1 clade phylogenetic analysis (Figure 6), it can be suggested that the rat ABCA7 is a fast evolving sequence. The rat ABCA7 gene exhibits long branches and it has not been clustered together to the rest of the ABCA7 orthologues (Figure 6), suggesting significant sequence divergence from the rest of the family. Although no disease has currently been associated to the human ABCA7 gene (Appendix II- Table 16), if in the future this occurs the rat would not be an appropriate animal model to use. The well supported phylogenetic analysis place ABCA1 and ABCA4 in a novel clade (Figure 6). Such a grouping can shed light on the common ancestry of these genes and infers potentially similar functionality.

The A2 clade is composed of two sub-clades A2a and A2b (Figure 7). The A2a sub-clade contains the genes ABCA3, ABCA14, ABCA15 and ABCA17. This sub-clade is in agreement with phylogenetic analyses that have also placed the ABCA14, ABCA15 and ABCA17 genes close to ABCA3 (Piehler *et al.*, 2006; Chen *et al.*, 2004). ABCA3 and ABCA17 are both placed in the position 16p13.3 of the human genome. The rodent specific ABCA14 and ABCA15 genes are located in the synthetic region of 16p12, suggesting a tandem gene duplication event subsequent to the rodent – primate divergence in mammalian evolution, potentially in adaptation to rodent fecundity (both genes being expressed in the testis) (Chen *et al.*, 2004). Further, the mouse ABCA15 gene exhibits long branches, suggesting a fast evolving sequence.

It is striking that all of the human genes present in the A2b sub-clade (ABCA5, ABCA6, ABCA8, ABCA9 and ABCA10) are present at the same locus, 17q24 (17q 24.3. 17q 24.3, 17q 24, 17q 24.2, 17q 24 respectively). In addition, they all exhibit a gene structure comprised of 38 exons, in contrast to the 50-52 exon structure of other ABCA genes (Albrecht and Viturro, 2007). Further, the divergence of such different structures occurred in early vertebrate evolution (Annilo *et al.*, 2006) and (Figure 7). Such commonality between the A2b sub-clade members supports their correct clustering within the ABC transporter superfamily. In reference to the primate-specific gene ABCA10 (Table 5), no

function or disease is currently associated to the gene itself, however if such a finding becomes apparent only primate model organisms should be considered for study. The duplication of ABCA8 occurred during the rodent lineage. Despite being well supported the position of ABCA8 duplicates in rodents (Figure 7) and the ABCB9 orthologues appear abherrant.

In general phylogenetic analysis of the ABCA genes reveals two distinct clades. These topologies are in overall agreement with previous heuristic analyses (Albrecht and Viturro, 2007).

ABCB genes

The ABCB genes are located within three clades, B1, B2 and B3 (Figures from 10 to 12).

Interestingly within the B1 clade (Figure 10) the two chicken ABCB1 and ABCB4 genes are grouped away from the mammalian clades suggesting two possibilities: separate mammalian and bird gene duplications from an initial ABCB1/4 precursor, or evidence of long branch attraction through the methods used.

The B2 clade (Figure 11) presents strong bootstrap support for the grouping of the ABCB2 and ABCB3 genes. These genes are half-transporters that form an heterodimer to transport peptides from the cytoplasm into the endoplasmatic reticulum (ER) that are presented as antigens by the class I HLA (Dean *et al.*, 2001b), are located at 6p21.3 and liked to Ankylosing spondylitis, insulin-dependent diabetes mellitus, and celiac disease (Appendix II - Table 17). Such evidence suggests tandem gene duplication prior to the mammalian radiation, where the genes have subsequently conserved much of the ancestral function and appear, to some extent, to be redundant in function.

The B3 clade (Figure 12) groups three half-transporters ABCB7, ABCB8 and ABCB10 that are located in the mitocondria of the cell (Dean *et al.*, 2001b).

ABCB7 is involved in the transport of heme from the mitochondria to the cytosol (Taketani *et al.*, 2003). Currently it is undetermined as to the function of the ABCB8 and ABCB10 gene, however based upon the function of their paralogues, it is possible that it exhibits a similar role.

ABCC genes

The ABCC genes are composed of two sub-clades, C1 (Figure 13) and C2 (Figure 14).
The C1 clade appears to have radiated during early vertebrate evolution, prior to the divergence of the avian and mammalian lineages. Within this clade ABCC8 and ABCC9 group together in one divergent clade and ABCC1, ABCC3, ABCC5 and ABCC6 group together (Figure 13).

Within the latter clade the ABCC1 and ABCC6 paralogues also lie adjacent at the genetic loci 16p.13.1, suggesting that tandem duplication has also characterised this family. Functionally ABCC6 is related to the recessive disorder Pseudoxanthoma elasticum (Plomp *et al.*, 2008), which causes skin in-elasticity and also affects the eye and cardiovascular system (Struk *et al.*, 1997). The grouping of ABCC1 and ABCC3 from a common ancestor is well established and appear to retain some shared function as they are both expressed within the placenta (Li *et al.*, 2007).

The absence of a chicken ABCC9 may be either due to the loss during avian evolution or because it has not been detected in the genome. However it is clear that ABCC9 is also an early vertebrate gene as teleost fish *(Danio rerio* also contain distinct ABCC8 and ABCC9 (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008)). Further the grouping of the C1 clade relates to the gene members function - both being potassium ion channel regulators (Bryan *et al.*, 2007). Further the C1 clade is of great relevance to medical research as both ABCC8 and ABCC9 have been implicated in diseases of relevance to drug discovery: ABCC8 has been related to Hyperinsulinemic hypoglycemia of infancy and non-insulin-dependent

diabetes mellitus type II (Table 18) and ABCC9 has been suggested to influence in myocardial infarction (Minoretti *et al.*, 2006).

The C2 clade (Figure 14) has robust statistical support for the evolutionary relationship of the genes ABCC13 and ABCC4 with CFTR and ABCC11 with ABCC12. CFTR acts as chloride channel and is associated with Cystic Fibrosis and congenital bilateral aplasia of the vas deferens (Appendix II- Table 18). The grouping of CFTR to ABCC4 and ABCC13 is a novel finding and thus, as these genes are not functionally elucidated, sheds light on the potential functions of such ABC transporters. Again, tandem duplication has characterised the evolution of the C2 clade as both ABCC11 and ABCC12 are located at 16q12.1, however the duplicates do not appear functionally conserved as ABCC11 is implicated in the type of earwax produced and ABCC12 has been related to breast cancer (Appendix II- Table 18).

ABCD genes

There is little statistical support for the topology of the ABCD phylogenetic tree (Figure 15), which may in part be due to the high sequence similarity between the sequences. Indeed, the initial Neighbor Joining tree (Digital Appendix) clusters the ABCD genes together with good support value. The ABCD transporters are involved in peroxisomal import of fatty acids and/or fatty acyl-CoAs in the organelle and are a family of half-transporters that require dimerisation to be functional (Liu *et al.*, 1999). Homo- and heterodimers are formed among the highly homologous ABCD1, ABCD2, and ABCD3 in vitro (Liu *et al.*, 1999) (Figure 15). However, heterodimers have not currently been detected with the more divergent ABCD4. ABCD genes are associated to peroxisomes disorders (PBD) such as Zellweger syndrome (ZS), which is correlated to two SNPs in ABCD3 (Gärtner *et al.*, 1992). Adrenoleukodystrophy (ALD) it is a rare inherited X-linked disorder caused by a disfunction in the ABCD1 gene. (Morita, 2007). ALD is a spectrum disease with varying severity. As no variation in ALD severity correlates with ABCD1 variants, additional genes may play a factor in this disease. Indeed the ABCD family may be

implicated. Studies in mouse with ABCD2 over-expression leads to correct the effects of the disease and ABCD3 only partially compensate (Braiterman *et al.*, 1998). Furthermore, ABCD4 expression correlates with the reduction in severity of ALD (Asheuer *et al.*, 2005), suggesting that ABCD4 is a principal modifier gene in ALD.

ABCE, ABCF and ABCG genes

ABCE, ABCF and ABCG are shown to be different from ABCB and ABCC genes but more similar among them (Figure 5). Detailed analysis of the GE clade of the ABC transporter superfamily (Figure 9) revealed a distinct separation of the two ABC subtypes as they for independent sub-clades. The separation of this clade supports the finding that ABCE and ABCF are structurally more homologues, preserving two NBFs but no TMs from the original ABC structure (Figure 4) (Dean *et al.*, 2001b). This finding demonstrates that ABCE, ABCF and ABCG genes form two distinct clades.

Although branching within ABCE mammalian orthologues could not be highly resolved and was subsequently collapsed (Figure 9), suggesting high similarity and potentially functionality in the different genes.
There are three ABCF genes each quite different in sequence (Figure 8 and Figure 14), and due to this divergence in sequence they may have non related functions.
There is strong support to group the ABCG1 and ABCG4 genes within the ABCG sub-clade (Figure 9). Indeed, ABCG4 was discovered due to its homology with ABCG1 (Velamakanni *et al.*, 2007), is suggested to heterodimerise with ABCG1 (Cserepes *et al.*, 2004). This idea is re-enforeced in the clade GE here presented, where it seems that ABCG4 was generated by ABCG1 gene duplication during mammalian evolution.

Different methods support the notion that the rodent specific gene ABCG3 is a duplicate of the ABCG2 gene. RACE sequencing generated the mouse ABCG3 sequence using the mouse ABCG2 in gene (Mickley *et al.*, 2001). However it

has been reported that even though the mouse ABCG3 gene is highly similar to ABCG2 it may not be functional (Mickley *et al.*, 2001). Within the rat both ABCG3 and an ABCG3-like gene have been identified (Appendix I – Table 15).

Even though within the ABCG sub-clade there is no robust bootstrap value to confirm the placement of ABCG5 and ABCG8 together, they are the only confirmed heterodimer within this family (Graf *et al.*, 2003). Indeed both ABCG5 and ABCG8 are implicated in the same diseases: Sterol accumulation, atherosclerosis and sitosterolemia (Appendix II- Table 19) and are present at the same genetic location 2p 21. Such findings suggest a tandem duplication during early vertebrate evolution, prior to the avian and mammalian divergences.

2.6 Conclusions

The phylogenetic analysis of the ABC transporter superfamily presented resolves the evolution of much of the gene family and reveals divergences in both structure and function, although it is not currently clear as to whether expression may have also subfunctionalised during this family's evolution. Further, tandem gene duplication appears to be a common factor in the evolution of the ABC transporter superfamily. By undertaking multiple phylogenetic analyses performed on multiple species using distinct methods we are able to assess common, and therefore strong, evolutionary signals and take into context those that are rarer, potentially due to biases in the algorithms. Such findings provide support for the use of multiple methodologies for robust phylogenetic methods.

Chapter 3. Phylogenetic analysis Black Box

This chapter describes the phylogenetic tool generated in order to perform robust phylogenetic analysis in an automated mode.

3.1 Introduction

The drug discovery process requires the analysis of protein sequences to address many different questions; for example to identify target sites, to reveal species-specific sequences for model organism selection and to assess the potential specificity of a drug based upon the homology of the target's paralogues. Each of these questions can be addressed using phylogenetic analysis. Phylogenetic analysis is low cost but impacts greatly at the gene to target phase, speeding up the process of drug discovery and diminishing its cost. However, robust phylogenetic analysis is not always performed because of time limitations or expertise in this field. For example, bioinformaticians may perform simplified phylogenetic analysis, based on one tree building method, which can often lead the user to over-interpret the tree or adopt false conclusions. Robust phylogenetic analyses, as performed in Chapter 2, can take a long time to perform manually and require expertise both in the methodologies used and the interpretation of multiple trees. The necessity of an automated alternative to run phylogenetic analysis is therefore essential to make robust phylogeny approachable and applicable to the drug discovery process.

3.2 Aims and objectives

The overall aim of this chapter is to describe the design and implementation of a user friendly programme to run robust phylogenetic analysis using multiple methods. This was done in order to generate a real, practicable, time saving alternative to manual phylogenetic analysis that can be integrated into the workflows of Computational Biologists at GlaxoSmithKline Pharmaceuticals (GSK).

The objectives of this project were to identify suitable phylogenetic methodologies and tools to infer the trees, generate a pipeline for phylogeny analysis and then to automate this process. In addition to the outputs of the phylogenetic methods, further objectives were to generate a guide report on tree interpretation, error reports and documentation.

3.3 Methods

The Phylogenetic analysis Black Box (PBB) program was developed and tested on GSK servers, IBM xSeries, running Red Hat Linux Enterprise Edition 4. The program was written in the PERL language (version 5.8.5). Two additional perl modules were installed from BioPerl version 1.5.1 (Stajich *et al., 2002*.) File::Copy enables the moving and copying of files in Perl language. Getopt::Long enables programs to take options from the user to modify its default behaviour. The Getopt:Long module also allows the options name to be longer than one character.

The following external software packages were already installed and were used in the PBB program:

- Readseq2 version 2.1.27 - sequence file conversion tool (Gilbert, 2001).
- PHYLIP version 3.6 - phylogenetic inference software (Felsenstein, 2005).
- TreePuzzle version 5.2 - phylogenetic inference software (Schmidt et al., 2002).
- MrBayes version 3.1.2 - phylogenetic inference software (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003).

3.4 Results

Four distinct phylogenetic analysis methods (Neighbor Joining, Maximum Parsimony, Maximum Likelihood and Bayesian) were selected in order to provide the user with phylogenetic trees from the most divergent phylogenetic algorithms. In order for the automation of phylogeny analyses, highly reliable phylogenetic software tools were required. The software and parameters used to manually perform the phylogenetic analysis in Chapter 2 were embedded into the automated pipeline. A workflow of the pipeline generated is presented in Figure 17.

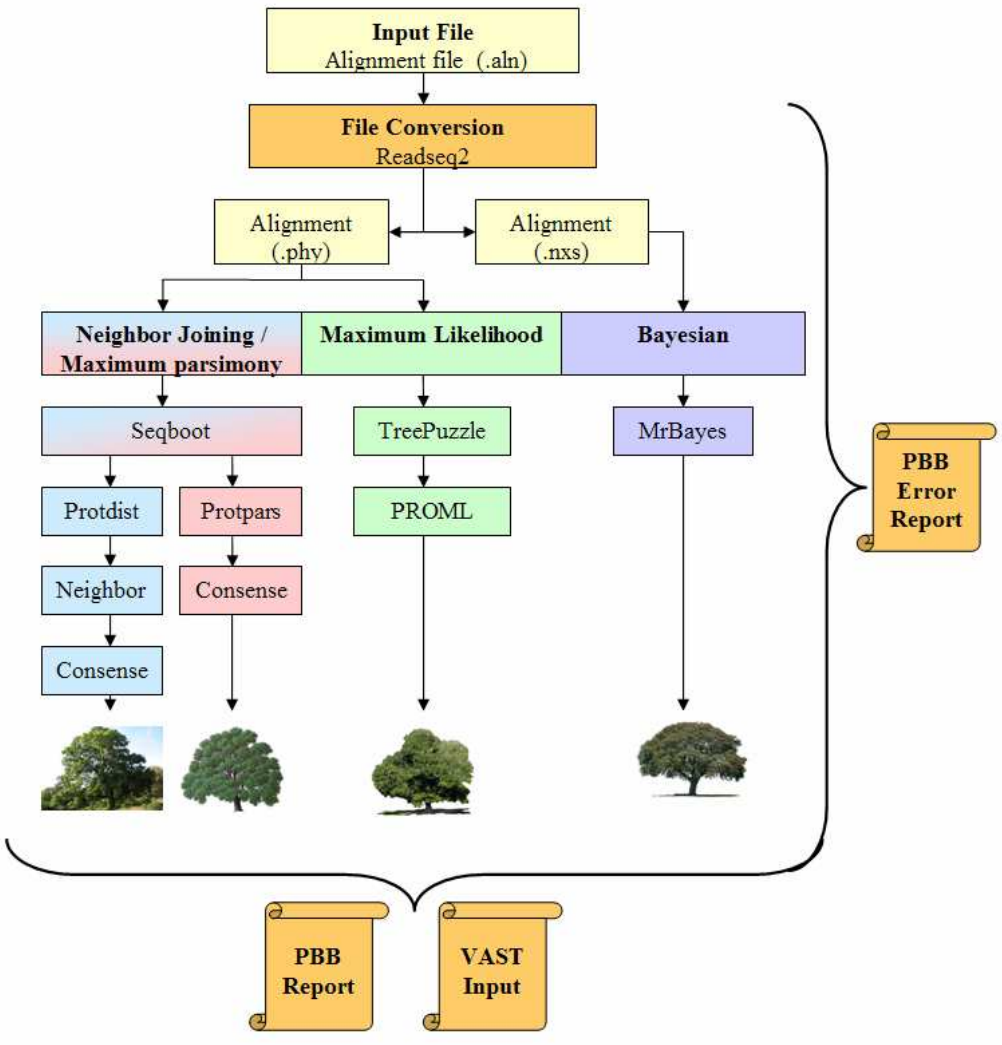


Figure 17. Phylogenetic Black Box (PBB) workflow

3.4.1 PBB Input

The PBB programme only accepts as an input aligned amino acid sequences in the CLUSTAL (.aln) format (Thompson et al., 1994) (Figure 18).

```
CLUSTAL 2.0.7 multiple sequence alignment


HS_ABCA4        FVRQIQLLLWKNWTLRKRQKIRFVVELVWPLSLFLVLIWLRNANPLFPNKAGMLPWLQGI
Cf_ABCA4        FARQIQLLLWKNWTLRKRQKIRFVVELVWPLSLFLILIWLRNINPLFPNKAGMLPWLQGM
Rn_ABCA4        FLRQIQLLLWKNWTLRKRQKIRFVVELVWPLSLFLVLIWLRNANPLFPNKAGMLPWLQGI
Mm_ABCA4        FLRQIQLLLWKNWTLRKRQKIRFVVELVWPLSLFLVLIWLRNANPLFPNKAGLLPWLQGI
Gg_ABCA4        FLRQVRLLLWKNWILRKRQKLRILVELIWPLSLFLALVWLRKANPLFPNKAGTLPWLQGI
Gg_ABCA2        FLRQVRLLLWKNWILRKRQKLRILVELIWPLSLFLALVWLRKANPLFPNKAGTLPWLQGI
Hs_ABCA1        CWPQLRLLLWKNLTFRRRQTCQLLLEVAWPLFIFLILISVRLSYPPFPNKAGTLPWVQGI
Cf_ABCA1        SWPQLRLLLWKNLTFRRRQTCQLLLEVAWPLFIFLILISVRLSYPPFPNKAGTLPWIQGI
Mm_ABCA1        CWPQLRLLLWKNLTFRRRQTCQLLLEVAWPLFIFLILISVRLSYPPFPNKAGTLPWVQGI
Gg_ABCA1        FWTQLGLLLWKNFTYRRRQTFQLLIEVAWPLFIFFILISVRLSYPPFPNKAGTLPWIQGI
HS_ABCA2        FLHQLQLLLWKNVTLKRRSPWVLAFEIFIPLVLFFILLGLRQKKPTFYTAAGILPVMQSL
Cf_ABCA2        FLHQLQLLLWKNVTLKRRSPWVLAFEIFIPLVLFFILLGLRQKKPTFYTAAGILPVMQSL
Rn_ABCA2        FLHQLQLLLWKNVTLKRRSPWVLAFEIFIPLVLFFILLGLRQKKPTFYTAAGILPVMQSL
Mm_ABCA2        FLHQLQLLLWKNVTLKRRSPWVLAFEIFIPLVLFFILLGLRQKKPTFYTAAGILPVMQSL
Rn_ABCA12       QFHQLRILVWKNWLGVKRQPLWTLVLILWPVIIFIILAITRTKFPPLAPRNGFFPFLQTL
Mm_ABCA12       QFHQLRILVWKNWLGVKRQPLWTLVLILWPVIIFIILAITRTKFPPLAPRNGFFPFLQTL
Rn_ABCA13       AGRQFQALLWKNWLCRLRHPVLSLAEFFWPCILFMILTVLRFQEPPLQARDGVLPFVQGL
Mm_ABCA13       AGRQFQALLWKNWICRLRHPVLSLAEFFWPCILFMILTVLRFQEPPLQARDGVLPFVQGL
Hs_ABCA13       AGCQFKALLWKNWLCRLRNPVLFLAEFFWPCILFVILTVLRFQEPPLQPRDGVIPFVQSL
Cf_ABCA13       METITPLLSLAPWSLQLRLETCRVLALPFALIYRARGAGLGAQILHLPVLTGGFEIIIRK
HsABCA7         SVQNHCPPCGLSPQESLGLALGQAQEPLHSLLEAAEDLAQELLALRLLQRPGPLELLSEA
Mm_ABCA7        QGSVTKLLEKILQRASLDPVLGQAQDSMRKFSDAIRDLAQELLTLPLLRRPGSLELVSEA
Rn_ABCA7        FCTQLMLLLWKNYTYRRRQPIQLVVELLWPLFLFFILVAVRHSHPPFPNKPGTVPWLQGL
Cf_ABCA12       NKSLKQICLLLNYINVISAGGSDNVTHVWHEGGQLSPSSLAAQLLILLNISADSPYIPYL
                                                          :   .   :

HS_ABCA4        FCNVNPCFQSPTPGEGIVSNYNNSILARVYRDFQEESQHLGRIWTELHILSQFMDTLRIR
Cf_ABCA4        FCNVNPCFQNPTPGEGIVSNYNNSILARVFRDFQEERQHFGHVWKEFQTLSRLMDTLRIR
Rn_ABCA4        FCNMNPCFQNPTPGEGTVSNYNNSILARLYRDFQEEVQRLGRVWTELRTLSQLMDTLPIR
Mm_ABCA4        FCNMNPCFQNPTPGEGTVSNYNNSILARVYRDFQEEVQHLGQVWAELRTLSQFMDTLQIR
Gg_ABCA4        FCNMNPCFRSPTRGEGVVSNYNNSILARVYRDAQEEIHDLGRVWEELIIMTQFMETMRIL
Gg_ABCA2        FCNMNPCFRSPTRGEGVVSNYNNSILARVYRDAQEEIHDLGRVWEELIIMTQFMETMRIL
Hs_ABCA1        ICNANPCFRYPTPGEGVVGNFNKSIVARLFSDARRKDTSMKDMRKVLRTLQQIKKSSKLQ
Cf_ABCA1        ICNANPCFRYPTPGEGVVGNFNKSIVSRLFSDAQRKDTSMKDIHEVLMTLQQVESFSKLQ
Mm_ABCA1        ICNANPCFRYPTPGEGVVGNFNKSIVSRLFSDAQRRDTSIKDMHKVLRMLRQIKHPNKLQ
Gg_ABCA1        ICNANPCFRYPTPGEGIVGNFNASIVSRLFSDAKRQDTSIKDVQKVLAKLRKLGNSSKLR
HS_ABCA2        CPDGQDEFGFLQYANSTVTQLLERLDRVVEEGNLFLGSELEALRQHLEALSAGPGTSSLD
Cf_ABCA2        CPDGQDEFGFLQYANSTVTQLLERLNRVVEEGNLFLGSELEALRQHLEALRLGPDTWSLG
Rn_ABCA2        CPDGQDEFGFLQYANSTVTQLLERLNRVVEESNLFLGSELEALHQRLEALSSGPGTWSLD
Mm_ABCA2        CPDGQDEFGFLQYANSTVTQLLERLHRVVEEGNLFLGSELEALRQRLEALSSGPGTWSLD
Rn_ABCA12       LCDTDKCKDTPYGPRLRRKGIDGTIFKESEVLKKSSNLSLQSTQVPERSHTSLATIPPRP
Mm_ABCA12       LCDTDKCKDTPYGPRLRRKGIDGPLFKESEILKKPSNLSLRSTQVPERSHTSLATVPPRP
Rn_ABCA13       LCNTGRCRNISFESSHRFRLPRFQTASDDRKVSSLQDLAEEILETMDKAKNLQKLWLKRS
Mm_ABCA13       LCNTGSCRNISFESSHHFRLSRFQTASDDRKVSSLQDLAEDIFETMDKAKNLQKLWLKRS
Hs_ABCA13       LCNTGRCRNFSYEGSHHFRLSRFQTAADPKKVNNLQDLAEEIHGMMDKAKNLKRLWVERS
Cf_ABCA13       RTYGANWLSRWSKRLLEFSSRFGATAGRREKVNDLQDLAEGIYEILDRAKILRELWAGGS
HsABCA7         LCSVRSTVGPSLNWYSDLMELVGQEPESALPDSSLLIGALDSHPLSRLLWRRLKPLILGK
Mm_ABCA7        LCSTKSPGGLSLNWYNQLNEFMGPEVAPALPDNSLFVGTLDDHPVSRLLWRRLKPLILGK
Rn_ABCA7        VCNVNSCFQHPTPGEGVLSNFKDSLISRLLADAHTVLGGHSTQDMLAALGKLIPVLRAVG
Cf_ABCA12       ACVRNVTDSLARGSQLRLLQSIISFKKSFLQNGFYVPEVLKSKLSQLRNVTELLCESTFD
```

Figure 18. Example of clustal format. Multiple sequence alignment from A2 clade from Chapter 2. (Digital Appendix).

3.4.2 PBB Tree Building-Methods

Neighbor Joining was performed in PHYLIP using the Seqboot, Protdist, Neighbor and Consense application (Felsenstein, 2005) (Figure 17). The Seqboot application generates 1000 replicates of the original input dataset, for each of which Protdist creates a distance matrix using the Jones Taylor Thornton algorithm (Jones et al., 1992). The Neighbor application subsequently infers a phylogenetic tree for each of the 1000 distance matrices generated. The input order of the sequences into Neighbor was randomised using a seed of 5 and jumbled 3 times. A consensus bootstrapped tree is generated from the 1000 replicates using the Consense application.

Maximum Parsimony was performed in PHYLIP using the Seqboot, Protpars and Consense applications (Felsenstein, 2005) (Figure 17). Much of the pipeline for Maximum Parsimony was replicated from the Neighbor Joining pipeline, however Protpars infers phylogenetic trees for each Seqboot dataset using a minimum evolution algorithm.

Maximum Likelihood was performed using two packages (Figure 17). TreePuzzle was used to generate an alpha value from the initial alignment (Schmidt et al., 2002). The alpha value was subsequently used by the PHYLIP ProML application (Felsenstein, 2005) in order to incorporate heterogenic rates of evolution across residues in the alignment into the Maximum Likelihood algorithm.

Bayesian analysis was performed using a Markov Chain Monte Carlo algorithm in the MrBayes application (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). 100,000 generations to the algorithm on six independent runs (chains) were used to infer the multiple Bayesian trees. 50% of these trees were sampled for subsequent consensus analysis. A consensus tree with clade credibility values was generated using the sumt command with exclusion (burn-in) of the first 1000 trees.

3.4.3 PBB Outputs

3.4.3.1 PBB Report

The PBB report is automatically produced during the PBB pipeline (Figure 17) and is a guideline in the interpretation of the resulting phylogenetic trees generated by the four methods. Therefore, the format of the PBB report has been designed to inform the user about the distinction of each algorithm performed and the reliability of the trees presented (Figure 19).

The PBB report gives the user a brief explanation of the model of evolution behind each method. The explanations are:

- Neighbor Joining: Distance based method, clusters similar sequences based in their similarity score.
- Maximum Parsimony: Searches for the minimum evolution tree.
- Maximum Likelihood: Searches for the most likely tree out from multiple generated.
- Bayesian: Searches for the most likely tree out from the original dataset based on posterior probability, bayesian statistics.

A tree reliability assessment is also produced for each tree generated. Different criteria for assessing the tree reliability were generated for the methods used. The final Neighbor Joining and Maximum Parsimony trees were both inferred using the PHYLIP Consense application (Figure 17) (Felsenstein, 2005) and thus the tree reliability interpretations for both methods are the same. Only clades presenting bootstrap values of 70% or more are classed as significantly credible, and clades with bootstrap values over 50% being considered as supported.

The tree topology is subsequently classed as one of the following:

- Strongly supported– more than 70% of the clades are credible.
- Well supported– 50% to 70% of the clades are supported.
- Supported - more than 50% of the clades are supported.
- Poorly supported – between 30% and 50% of the clades are supported.
- Not supported – less than 30% of the clades are supported.

Maximum Likelihood is not a bootstrap method. Therefore, PBB cannot provide a reliability assessment. The user is informed that Maximum Likelihood produces the most statistically likely tree and after performs better than Neighbor Joining and Maximum Parsimony (Kuhner and Felsenstein, 1994).

Based on analyses in Chapter 2 Bayesian phylogenetic tree inference has been deemed over confident. Therefore clades presenting clade credibility values (equivalent to bootstraps) of 90% or more are classed as significantly credible, and clades with clade credibility values over 70% being considered as supported. The tree topology is subsequently classed as one of the following:

- Strongly supported– more than 70% of the clades are credible.
- Well supported– more than 70% of the clades are supported.
- Supported - more than 50% of the clades are supported.
- Poorly supported – between 30% and 50% of the clades are supported.
- Not supported – less than 30% of the clades are supported.

Phylogenetic Black Box version 1.0
Input file: infile
Date:

---------Bayes -------
Method explanation

PBB tree reliability

---------Neighbor Joining -------
Method explanation

PBB tree reliability

---------Maximum Parsimony -------
Method explanation

PBB tree reliability

---------Maximum Likelihood -------
Method explanation

PBB tree reliability

Contact information:
Phylogenetic and programming experts

Figure 19. Stylised view of PBB report

3.4.3.2 PBB Error Report

The Phylogenetic Black Box also generates an error report in case any error is produced during the pipeline (Figure 20).The format of the PBB error report is presented in Figure 21 and reflects that seen in the PBB report itself. The original error messages of the programs where the errors have been produced are printed into the file together with by an explanation what this error means and finally tips or hints of how to solve the problem are printed to the user. Furthermore, in order to aid the user on understanding where in the pipeline the error occurred and how subsequent outputs may be affected, the workflow is printed (Figure 20).

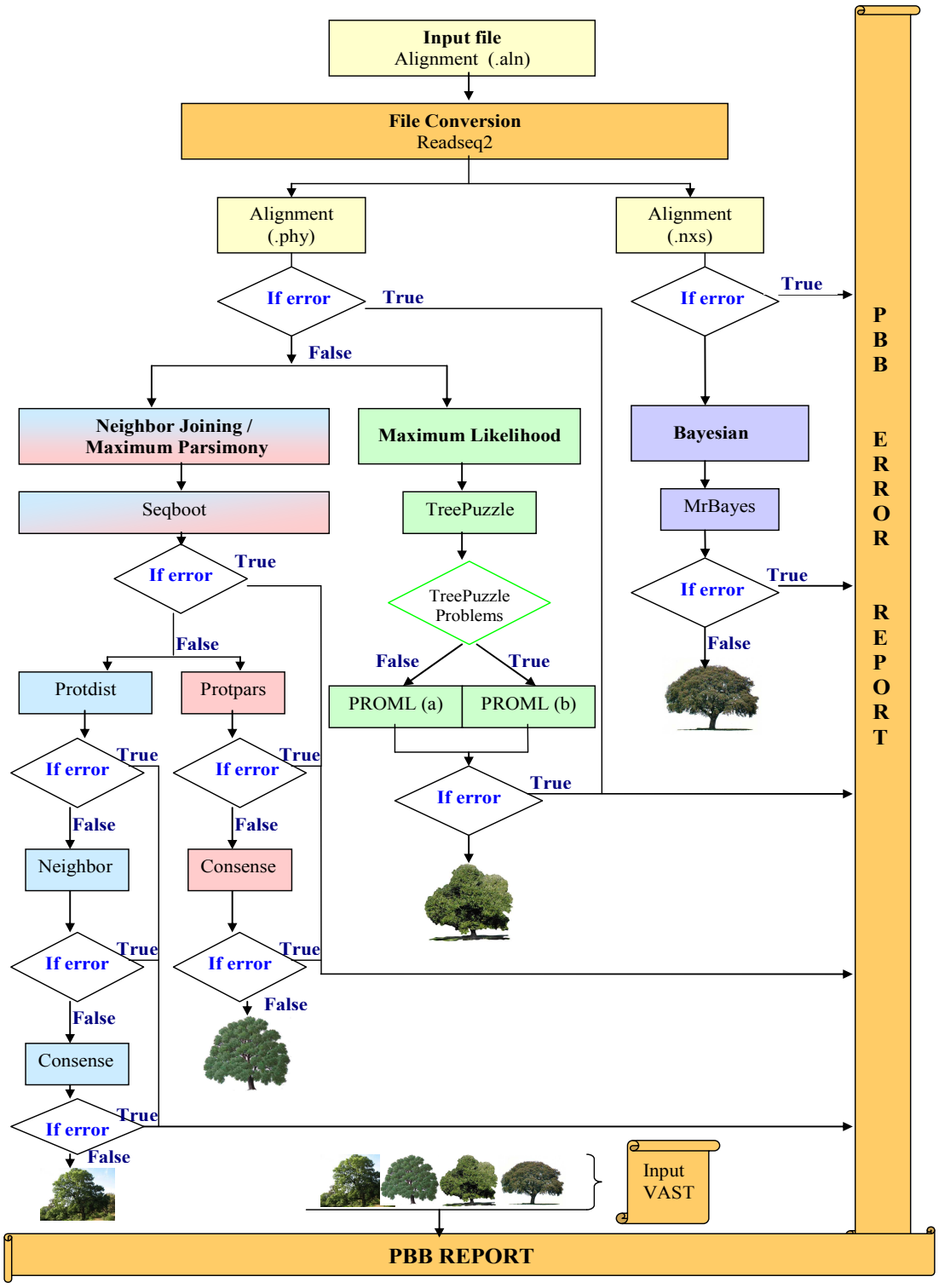


Figure 20. Error checking over the pipeline

Phylogenetic Black Box version 1.0

Input file: infile

Date:

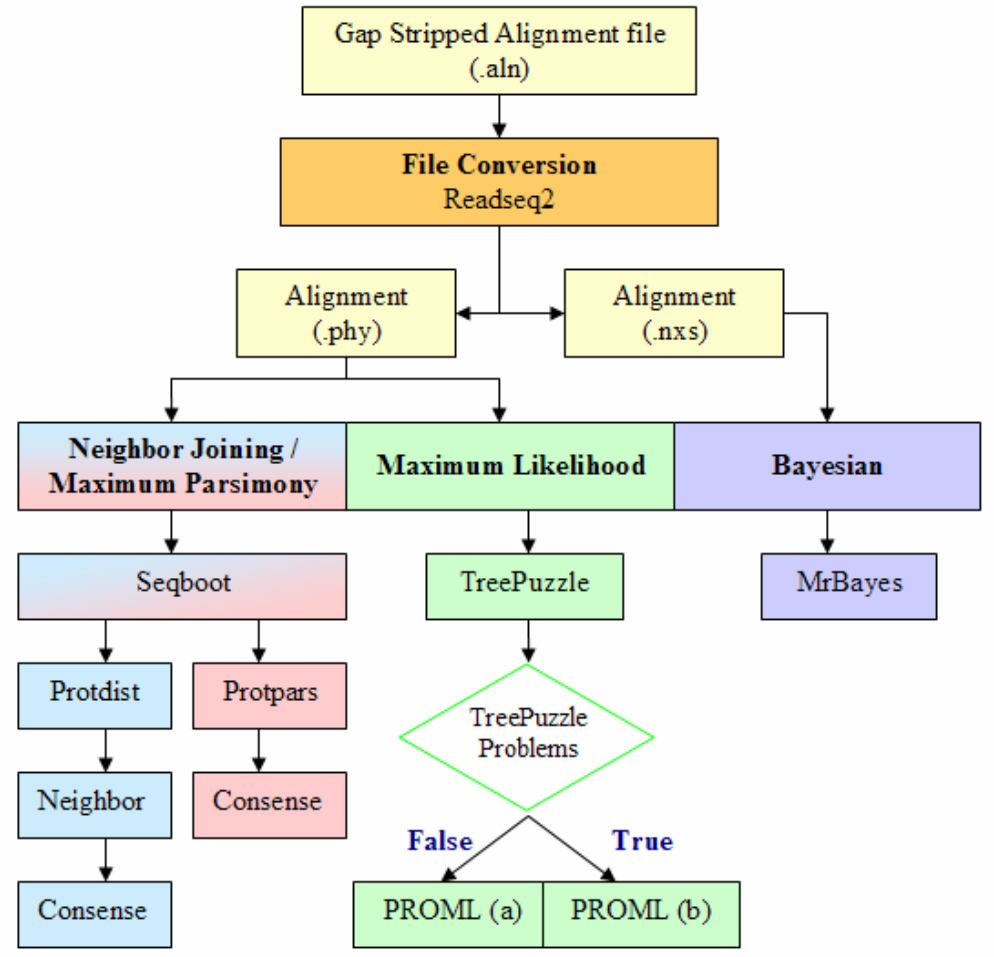
Errors detected:

Pipeline Program Errors

PBB Error interpretation

Tips to solve the errors

Pipeline workflow



Contact Information: PBB maintenance

Figure 21. Stylised view of PBB Error Report

3.4.3.3 Phylogenetic trees

In order to generate the four final trees, the PBB pipeline produces 40 documents. From those, by default, only the trees are preserved in order that they can be subsequently individually viewed by the user in tree visualisation softwares such as NJplot (Perrière and Gouy, 1996).

3.4.3.4 VAST Input

An input to the tree viewing software VAST (Mackey, 2007) is also automatically generated during the PBB workflow. The input generated has the format presented in Figure 22.

Figure 22. Stylised view of VAST Input format.

3.4.4 Software design

The software was designed in order to fulfil the objectives mentioned in section 55. Using the PBB workflow (Figure 17) the software can be divided in five sections discussed below. The code for the software is presented in Digital Appendix.

3.4.4.1 Input command

To execute PBB on the command line, the user can input the following command:

PBB –infile *infile*

The infile is received as an argument and refers to the amino acid alignment PBB is to perform the different phylogenies on. Further one or more additional options can be called:

PBB –infile *infile* -a –out -help

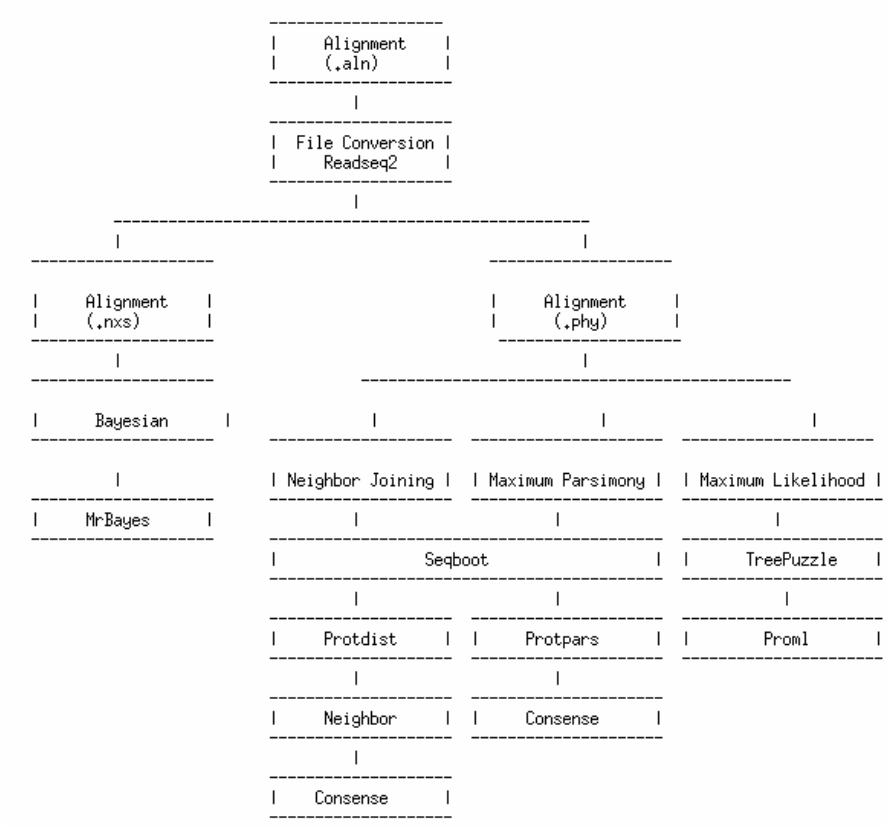
The PBB workflow generates 40 files, from which, 32 intermediate files are deleted by default, providing to the user the eight relevant outputs as described in Section 3.4.3. By adding the option "-a" all files are retained in the output. Automatically the PBB program creates a folder where to keep all the files created during the workflow. This folder, by default, is named by concatenating "PBB" and "*infile"*. By using the option "-out" the user can change the final base working directory, instead of the default of the current working directory. Finally, the "-help" option prompts information about the program to the user as displayed in Figure 23.

```
Description:
Phylogenetic Black Box (PBB) is a phylogenetic
pipeline. From an initial amino acid alignment it
produces four different phylogenetic trees.


Input files
Amino acid alignment
Format : clustal (.aln)


Tree building methods:
 -Neighbor Joining: Distance based method.
  Tree constructed as successive clustering of
        lineages. Saitou & Nei (1971).
 -Maximum Parsimony:
        Minimum evolution tree."The simplest the the logical"
 -Maximum Likelihood:
  Most likely tree out of a random set generated.
 -Bayesian:
        Application of Bayes theorem: probability of a tree based
        on the observations of multiple trees performed.


Pipeline workflow:
```




```
Usage:
 PBB_v21.pl <AA alignment> [options]

Options:
--help              this
--a                 keep all output files
--out               choose output directory
  Always give full path in options

Example:
  perl PBB_v21.pl infile -a -out='/home/mudid/mydirectory'
```

Figure 23. PBB help information as prompted to the user when the –help command is executed

3.4.4.2 File conversion

The format of the alignment provided as an input into PBB (.aln) is not adequate for starting the different tree-building methods. Therefore, the program calls the application Readseq2 to transform the CLUSTAL (.aln) format to the PHYLIP (.phy) and nexus (.nxs) formats for PHYLIP, TreePuzzle and MrBayes (Figure *17*).

3.4.4.3 Pipelining the tree building methods

The pipelining of the tree building methods for Neighbor Joining, Maximum Parsimony, Maximum Likelihood and Bayesian were performed following the methodologies, softwares and parameters presented in Section 3.3.

The applications called in the pipeline need parameters defined prior to execution. In order to address this problem, the PBB program generates a file called “inputparametersprogramname” on the fly for each application. In this file the parameters to be used by the application are printed. The file is then called as an input to the application, which subsequently reads in the parameters and executes according to them.

The applications called in the pipeline also require the appropriate input file to execute on. Applications provided by the PHYLIP package search by default an input file called “infile” and output their results to files called “outfile” and “outtree”, the latter only being generated in applications that generate trees. PBB breaks with this paradigm and the output files are renamed following PBB nomenclature. The input files are provided as a parameter. TreePuzzle and MrBayes accept the input file like a parameter as default.

3.4.4.4 Error checking

Errors can occur during the running of the PBB programme. The pipeline incorporating error checking is presented in Figure 21. The main errors are addressed below.

Infile format not appropriate

Detection of the infile format is determined by checking the Readseq2 output. Readseq2 detects 25 different biosequence formats (Gilbert, 2001). The CLUSTAL format is number 22. Therefore, if Readseq does execute correctly, the input file has a different format number to that of CLUSTAL and an error is reported and the PBB pipeline terminates.

Files deleted while the program is running

Each time an application is called the existence of the input file is checked (Figure 21). The lack of the appropriate input file for the application results in the failure of one or more phylogenies and an error message is reported.

Applications not installed

PBB calls different external applications that if absent result in the termination of the current PBB application and an error message is reported

User terminating the PBB program while it is running

If the user wants to terminate PBB while it is running using “Ctrl + C”, PBB stops the current application and tries to continue with the next one. Pressing “Ctrl +C” may cause the absence of one or more tree files as an output.

Errors or warnings prompted by the external programs

Different applications called by the PBB program may emit diverse error messages to the user. These messages are preserved in the PBB Error Report for the user’s awareness. The error message format varies depending on the application. By convention, Readseq2 and PHYLIP start the error messages by “Error”. Besides explicit errors, PBB would consider as an error any undesirable

output parameters in Readseq2. The MrBayes application emits two kinds of error messages. Severe errors start by “Error” followed by the string “Could not” and an indication of where the error is explained in further detail. Non-severe errors present the word “warning” and the next two lines explain the causes of the warning. TreePuzzle errors are not checked. If the alpha value is not produced the next program, ProML, does not input it as a parameter.

Unfinished run of applications

The applications that perform the inference of the distinct phylogenies, Neighbor, Protpars and ProML, run computationally demanding algorithms. The input of a high number of sequences with a high number of characters to compare can cause the non-complete execution of these applications.

Neighbor and Protpars applications infer phylogeny of 1000 dataset replicates generated by Seqboot. The incomplete execution of the applications results in an output file with less than 1000 trees, with the last tree inference unfinished. This file is not a valid input to Consense, the final application of both methodologies, and as a consequence, no phylogenetic tree would be generated in the Neighbor Joining and/ or Maximum Parsimony parts of the pipeline. To handle such situations PBB checks for the correct ending of the files generated by the Neighbor and Protpars applications and trims the final incomplete tree in case of its occurrence. The incorporation of the correction allows PBB to retrieve an accurate phylogenetic tree from both methodologies, even though the number of replicate data sets considered is less than 1000. The number of replicates used in either would be shown in the PBB report with the corresponding phylogenetic tree.

The ProML algorithm is much more demanding than Neighbor or Protpars. It takes a long time to run even with only a few sequences to infer phylogeny from. As it only generates one tree as an output, if the execution of the application is not completed the output cannot be retrieved or presented to the user. To mitigate the run-time PBB performs the Maximum Likelihood

methodology at the end of the pipeline. In case ProML execution does not complete soon enough, the user can press “Ctrl+C” for PBB to generate the report where the Maximum Likelihood tree would not be present.

3.4.4.5 Generation of outputs

PBB Report

The PBB Report format (Figure 19) is generated on the fly by the program. As soon as a phylogenetic tree is output, the PBB workflow writes the phylogenetic methodology name, a general explanation of the methods, the phylogenetic tree and the PBB tree reliability calculations to a temporary file. The final PBB Report is generated when the entire workflow has been performed, using the information from the temporary file and adding the heading and the contact information (Figure 19). The initial file is deleted by default.

PBB Error Report

The PBB Error Report is generated in a similar way to the PBB Report. An initial error report is created on the fly and the error information is appended to it each time an error is detected. Once the final workflow has been run, the PBB Error Report is created containing the same information as the temporary file but adding the heading of the document, the workflow schema and the contact information (Figure 21). The initial file is deleted by default.

Phylogenetic trees

Phylogenetic trees are produced automatically by the final applications called for each methodology. These are Consense for Neighbor Joining and Maximum Parsimony, ProML for Maximum Likelihood and MrBayes for Bayesian analysis. The names of the files containing the phylogenetic trees are presented in Table 9. The files allow the user to observe the trees in different visualisation softwares.

Table 9. File names of the phylogenetic trees presented for each tree-building method.

Method	File nomenclature
Neighbor Joining	alignmentfilename.aln.phy.Seqboot.Protdist.neighbor_tree.Consense_outfile
Maximum Parsimony	alignmentfilename.aln.phy.Seqboot.Propars_tree.Consense_outfile
Maximum Likelihood	alignmentfilename.aln.phy.maximum_likelihood_outfile
Bayesian	alignmentfilename.aln.nxs.con

VAST input

The creation of the VAST input is performed once the whole PBB pipeline has been run as it requires the final outputs of each methodology (Table 10). The final format can be seen in (Figure 22).

Table 10. File names of the files required to create Input VAST.

Method	File nomenclature
Alignment	alignment.aln.nxs
Neighbor Joining tree	alignmentfilename.aln.phy.Seqboot.Protdist.neighbor_tree.Consense_outtree
Maximum Parsimony tree	alignmentfilename.aln.phy.Seqboot.Propars_tree.Consense_outtree
Maximum Likelihood tree	alignmentfilename.aln.phy.maximum_likelihood_outtree
Bayesian tree	alignmentfilename.aln.nxs.con

3.5 Usage

In order to generate the PBB workflow and the different reports, PBB was tested using the ABC clade analyses performed in Chapter 2. In addition, further alignments of interest were provided by members of the Computational Biology department at GSK (Table 11). The first four analyses show the typical alignments inputted by the computation biology users and takes up to three hours on the servers available. However, the PBB run-time varies depending on the amount of sequences and sites to compare in the alignment provided. PBB was also tested with a very large alignment (Table 11). In this case the programme run overnight.

Table 11. PBB analysis performance by alignments provided by members of the Computational Biology department at GSK.

Alignment	Sequences	Characters	Time (h)
A	17	124	0:50
B	13	425	1:15
C	16	451	2:30
D	12	837	3:20
E	94	155	15:17

3.6 Server

The PBB tool was implemented on a server at GSK, where all Computational Biology users have access. From this location PBB is available to any user of the server by using the commands presented at section 3.4.4.1.

3.7 Documentation

PBB has been documented to make it approachable to the users and facilitate its further development. Basic information about the program can be found by inputting the help option in the command line. A basic tutorial of the program is provided at the wiki page of the Computational Biology department at GSK, where, this chapter and the whole thesis, are linked for more detailed information on PBB and phylogenetics. Finally, the scripts are available at the server. They have been documented in detail in order to facilitate the implementation of future improvements in the current version of the program.

3.8 Discussion

Phylogenetic analysis Black Box (PBB) is a pipeline of distinct phylogenetic tree building methods: Neighbor Joining, Maximum Parsimony, Maximum Likelihood and Bayesian. Its main function is to generate, from an initial amino acid alignment, four distinct phylogenetic trees. Furthermore, PBB helps in the understanding of the results for users with no expertise in phylogenetics. The general workflow of PBB as seen from a user's point of view is depicted in Figure 24.

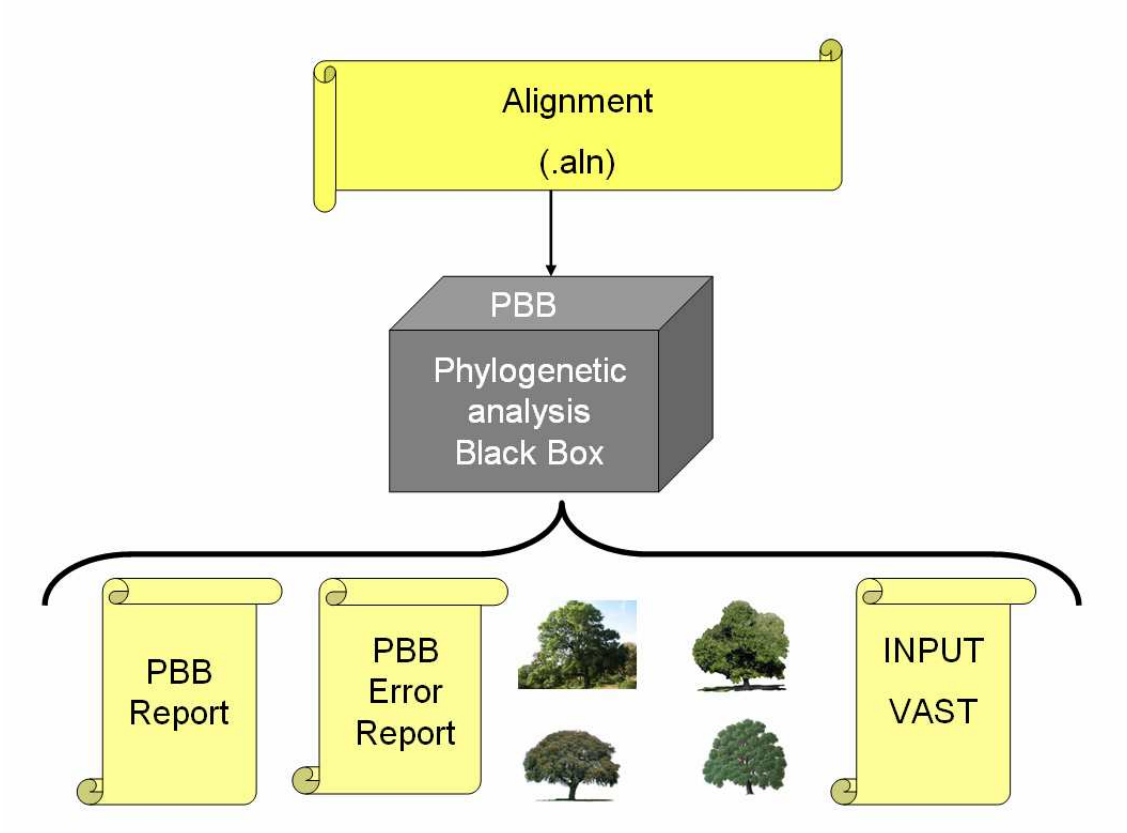


Figure 24. PBB user-view workflow.

In order to simplify the analysis, the PBB programme only accepts as an input aligned amino acid sequences in the CLUSTAL (.aln) format (Thompson et al., 1994). The user can generate the alignment in the software of their preference (e.g CLUSTALW, CLUSTALX, MUSCLE, T-Coffee) and subsequently convert the alignment into the CLUSTAL format prior to PBB analysis. Furthermore, this

option allows the PBB input to become relatively future-proofed for as newer methodologies for alignment are produced, PBB can still be used as long as the alignment is converted into this standard CLUSTAL format. In addition, the user is recommended to provide an alignment of more than 50 sites to infer a robust analysis.

The five outputs of the PBB programme enable a non-expert user to quickly perform robust phylogenetic analysis. The tree inference report enables a non-expert user to understand the complexities in phylogenetic analysis and interpret the trees correctly. The error report informs the user of any errors that happened during the runtime process. The four trees are produced in order for the user to study each tree individually. The VAST input allows the user to observe the alignment at the same time as the tree and quick visualise the four trees using the VAST software (Mackey, 2007).

PBB documentation guides the user in the execution of the program and the understanding of phylogenetics. Furthermore, documentation about the coding facilitates programmers to improve and expand PBB functionality.

3.8.1 Limitations

During the testing of the programme two potential limitations were observed. As it has been reported in Section 3.5, if the input alignment contains large amount of sequences or long sequences it is likely that a Maximum Likelihood tree will not output. Therefore, the user would obtain only three phylogenetic trees from the PBB. Secondly, the input of very divergent sequences may not produce a convergent Bayesian tree. In such cases a warning would be reported to the PBB Error Report. The limitations presented here are the same limitations as in a manual phylogenetic analysis.

3.8.2 Expansion

Expansion into VAST

An additional aim of the project was to integrate the PBB with other tools currently used by Computational Biologists at GSK. One such application is VAST, web-based tree viewing software produced by Aaron Mackey at GSK (Mackey, 2007). In order to link these two tools, an additional output of the PBB called “Input VAST” was generated that can be directly entered into the VAST application, in order to view and edit the alignment and trees produced by the PBB. An example of the format of the PBB output when viewed in the VAST application is presented in Figure 25.

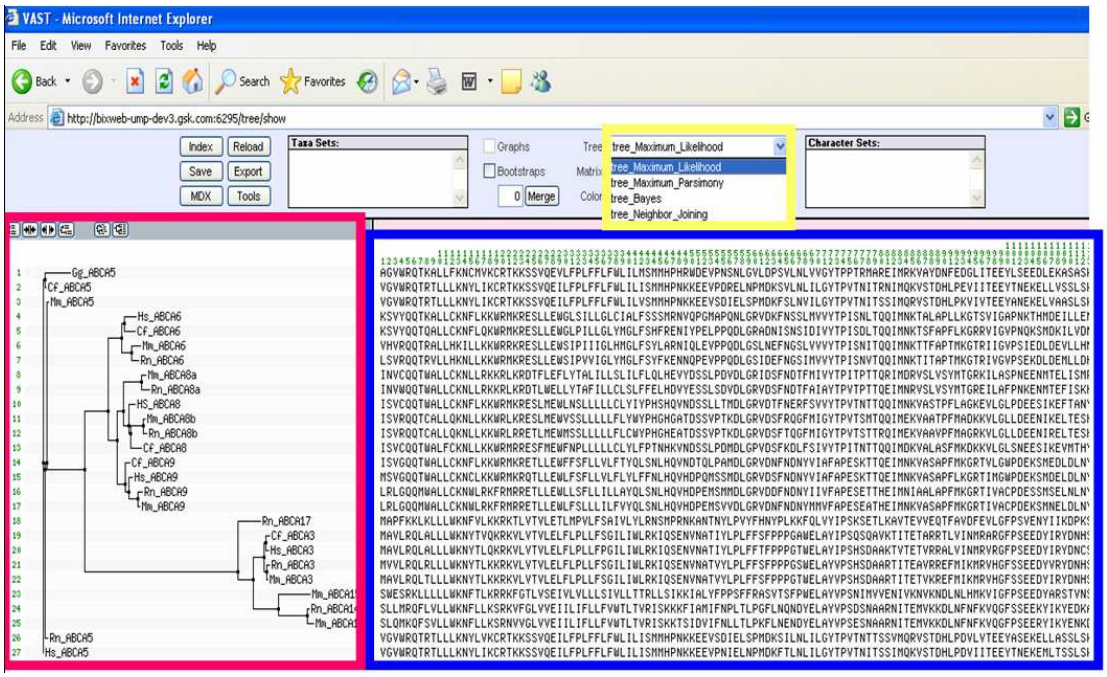


Figure 25. Example of PBB output viewed in VAST.

Blue square: *Visualisation of the entire alignment.* ***Red square:*** *on the left of the software the user can visualise and edit the phylogenetic tree.* ***Yellow square:*** *On the top the user has the option of changing the tree to visualise.*

A further step would be to add PBB as an application in VAST itself. Having PBB as a tool in VAST would allow the VAST users to perform the four phylogenies while they are observing the alignments and enable non-unix or unwilling command-line users to employ PBB.

Parallelising the pipeline

PBB is already a significantly time-saving alternative to manual analysis (Section 3.5). Multi-threading the process would improve PBB time performance. However, the PBB pipeline would still be limited by the longest running single applications, usually MrBayes or ProML. Improvement in either, perhaps through multithreading, would greatly impact on long alignments.

Timing Maximum Likelihood algorithm

Maximum Likelihood trees are not always obtained (Sections 3.4.4.4 and 3.8.1). In order to avoid the need for the user to press “Ctrl + C” when the Maximum Likelihood application does not complete, after an arbitrary (possibly insufficient) delay, an automatic mechanism should be incorporated for timing the application run and to stop it when a certain, reasonable, amount of time has passed if the output has not been generated.

Nucleotide Phylogenetic analysis Black Box

At the moment PBB only performs phylonegenetic analysis when the input is an amino acid alignment. However, nucleotide phylogenetic analyses are also relevant in drug discovery. In order to create a nucleotide phylogenetic alignment black box a similar pipeline to the current PBB could be programmed but calling different applications and inputting their respective parameters. This resulting program can stand alone as Nucleotide PBB or can be merged with the current amino acid PBB “AA PBB” into a single PBB where amino acid or nucleotide alignments can be inputed and the phylogenies inferred.

3.9 Conclusions

The Phylogenetic analysis Black Box program developed in this chapter performs all the required tasks correctly. Furthermore, PBB has been demonstrated to be a time saving alternative to the equivalent manual phylogenetic analysis (Section 3.5) and hence is a valuable tool in the drug discovery process. PBB could be further improved as has been discussed in (Section 3.8.2). However, simply having a phylogenetic analysis black box for amino acid sequences significantly increases the amount of robust phylogenetic analysis performed in the Computational Biology department at GSK and speeds up the whole drug discovery process. Furthermore, by providing documents to train the user (basic help, wiki page, scripts-Digital Appendix) and to aid the non-expert users in tree interpretation (PBB Report) the PBB programme makes phylogeny more approachable to non expert users and fully incorporates all of GSK's needs.

Chapter 4. General Discussion

This thesis guides the user through molecular phylogenetics explaining the fundamental principals of this bioinformatics method (Chapter 1) and providing a detailed example of robust phylogenetic anlaysis and result interpretation (Chapter 2). From both chapters it is clear that phylogenetics has relevance in the drug discovery process, mainly in the drug targeting phase and for identifying appropriate animal models.

As phylogenetics is a purely theoretical tool to infer the evolutionary history of genes, species etc no method can guarantee to provide the correct evolutionary tree. However, from a phylogenetic tree a wealth of diverse information can be inferred: evolutionary divergence, prediction of changes in protein structure, functionality, expression etc. Indeed in a gene family study, by combining current understanding of some of the paralogues, functions may be proposed for gene family members with no current identified function. Combining and overlaying phylogenetic analyses with literature is a most common procedure in phylogenetics publications.

Chapter 2 demonstrates the necessity of performing more than one method in phylogenetic analyses as no method guarantees to obtain the correct evolutionary relationship. This issue has been experimentally observed when in the ABC subfamily clades (Figure 6 to 15) distinct methods were giving high bootstrap support to different branches. This finding re-enforces the notion that not all the methods infer the same evolutionary relationship. However, it is clear that the most closely related sequences tend to be highly supported for the four distinct methods utilised. In addition as the methods used, Neighbor Joining , Maximum Parsimony, Maximum Likelihood and Bayesian analysis, infer phylogenies based on very different mathematical models, if a relationship is supported from each method it is very likely to be true. In contrast, one highly supported relationship from only one method it is not necessarily false, it is

feasible that only one method is able to depict that relationship because as it has been explained in Chapter 1 each method has its own biases, however it is less likely. Further, it is possible for a relationship depicted by all four methods to be incorrect – for example in the ABC transporter A2 clade (Figure 7) the ABCA8 rodent duplicates and the ABCA9 clade show good statistical support from each method, however, it does not make evolutionary sense as ABCA8 rodent duplicates are more likely to share common ancestry with their ABCA8 orthologues as opposed to the ABCA9 paralogues. Even though incorrect topologies from multiple methods are rare, such an event demonstrates the necessity to compare the results with more than one method and with the literature available to suggest possible functions and conclusions of the sequences being analysed. This procedure has been followed in Chapter 2, and mirrored in the automated phylogenetic anlaysis application Phylogenetic Black Box (PBB) by giving a calculation of the statistical support of each tree provided (Chapter 3).

PBB is a bioinformatic tool to run robust phylogenetic analysis under predetermined parameters (Chapter 3). Its creation was the main aim of the thesis, in order to address the requirement of robust automated phylogenetic analyses at GlaxoSmithKline Pharmaceuticals (GSK). It was clear that in the drug discovery industry the lack of time appears to be the main reason why multiple methods have previously not been utilised to a great extent when assessing the potentiality of a drug target. Further, to perform such analyses, the user is required to have an expert knowledge in the different phylogenetic methods to implement. Such problems are reduced with PBB, which runs robust phylogenetic analysis using the Neighbor Joining, Maximum Parsimony, Maximum Likelihood and Bayesian methodologies and has demonstrated to be a time saving alternative to manual analysis (Table 11), as that performed in Chapter 2.

Another issue arising during the course of the thesis is that if a phylogenetic analysis uses multiple methods, a consensus tree is presented. At the initiation

of the PBB project the creation of a consensus tree from the four trees was considered. However the notion was subsequently discarded as a consensus tree prevents the user from understanding the phylogenetic data and potentially biasing them towards a position of thinking that the consensus tree is the true tree, which is not the case. With the aim of helping users in phylogenetic interpretation, PBB provides the PBB report where a calculation of the global tree support is performed and brief descriptions of the methods performed are reported (Digital Appendix). In addition, as PBB is for general use the parameters selected are those considered for standard phylogenetic analysis performed at GSK. The methods and parameters have been selected according to GSK phylogenetics experts in the Computational Biology department. For users with knowledge in phylogenetics, one can change the parameters used in the PBB application in order to perform a phylogenetic analysis in accordance to the type of data they have or the method selected. Further, GSK Computational Biologists can copy and modify the PBB script enabling the PBB to be integrated in phylogenetic analysis at GSK when new methods arise.

Chapter 5. General Conclusion

This thesis provides phylogenetic guidance for those with no previous expertise in this topic. The example given by the ABC transporters superfamily represents a complex but interesting phylogenetic analysis, where trees have been interpreted and evaluated using the different methodologies. Further, such analysis has elucidated the true relationships between this family and reveals novel insights into the evolution and diversification of ABC transporter function. Finally, PBB has been generated to automate the time-costly process of manual phylogenetic analysis and to guide non-expert users through the complexities of phylogenetics. PBB is currently available at the Computational Biology department at GSK, where many researchers are utilising it as a tool to run phylogenies that otherwise would have been limited to only one phylogenetic method.

5.1 Further work

Further work in the Phylogenetic analysis Black Box can be done. As it has been mentioned in Chapter 3 (Section 3.8.2) expansions to the tool can be performed in two distinct directions – improving PBB resolution speed and making it more user friendly. For the last feedback from users would be a great source of information. Once the current PBB is of general use the next step would be to generate a PBB tool for nucleotide sequences. The main aim is to provide at the Computational Biology department at GSK with a tool to peform rapid reliable robust phylogentic analysis in both kinds of sequences – amino acid or nucleotide ones.

References

Albrecht, C. and Viturro, E. (2007), "The ABCA subfamily - Gene and protein structures, functions and associated hereditary diseases", *Pflugers Archiv European Journal of Physiology,* vol. 453, no. 5, pp. 581-589.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990), "Basic local alignment search tool", *Journal of Molecular Biology,* vol. 215, no. 3, pp. 403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997), "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs", *Nucleic Acids Research,* vol. 25, no. 17, pp. 3389-3402.

Annilo, T., Chen, Z. -., Shulenin, S., Costantino, J., Thomas, L., Lou, H., Stefanov, S. and Dean, M. (2006), "Evolution of the vertebrate ABC gene family: Analysis of gene birth and death", *Genomics,* vol. 88, no. 1, pp. 1-11.

Annilo, T., Chen, Z. -., Shulenin, S. and Dean, M. (2003), "Evolutionary analysis of a cluster of ATP-binding cassette (ABC) genes", *Mammalian Genome,* vol. 14, no. 1, pp. 7-20.

Annilo, T. and Dean, M. (2004), "Degeneration of an ATP-binding cassette transporter gene, ABCC13, in different mammalian lineages", *Genomics,* vol. 84, no. 1, pp. 34-46.

Asheuer, M., Bieche, I., Laurendeau, I., Moser, A., Hainque, B., Vidaud, M. and Aubourg, P. (2005), "Decreased expression of ABCD4 and BG1 genes early in the pathogenesis of X-linked adrenoleukodystrophy", *Human Molecular Genetics,* vol. 14, no. 10, pp. 1293-1303.

Baldauf, S. L. (2003), "Phylogeny for the faint of heart: A tutorial", *Trends in Genetics,* vol. 19, no. 6, pp. 345-351.

Baxevanis, A. D. and Ouellette, B. F. F. (2004), *Bioinformatics : a practical guide to the analysis of genes and proteins,* 3rd ed, Wiley-Interscience, Hoboken, N.J.

Bouige, P. (2008)*, Institut Pasteur- Unit of Molecular Programming and Genetic Toxicology (UMPGT)*, available at: http://www.pasteur.fr/recherche/unites/pmtg/about.html (accessed 20th of August 2008).

Braiterman, L. T., Zheng, S., Watkins, P. A., Geraghty, M. T., Johnson, G., McGuinness, M. C., Moser, A. B. and Smith, K. D. (1998), "Suppression of peroxisomal membrane protein defects by peroxisomal ATP binding cassette (ABC) proteins", *Human Molecular Genetics,* vol. 7, no. 2, pp. 239-247.

Bryant, D. (2005), "On the uniqueness of the selection criterion in neighbor-joining", *Journal of Classification,* vol. 22, no. 1, pp. 3-15.

Bryan, J., Muñoz, A., Zhang, X., Du□fer, M., Drews, G., Krippeit-Drews, P. and Aguilar-Bryan, L. (2007), "ABCC8 and ABCC9: ABC transporters that regulate K+ channels", *Pflugers Archiv European Journal of Physiology,* vol. 453, no. 5, pp. 703-718.

Chakrabarti, S., Lanczycki, C. J., Panchenko, A. R., Przytycka, T. M., Thiessen, P. A. and Bryant, S. H. (2006), "Refining multiple sequence alignments with conserved core regions", *Nucleic Acids Research,* vol. 34, no. 9, pp. 2598-2606.

Chen, Z. -., Annilo, T., Shulenin, S. and Dean, M. (2004), "Three ATP-binding cassette transporter genes, Abca14, Abca15, and Abca16, form a cluster on mouse chromosome 7F3", *Mammalian Genome,* vol. 15, no. 5, pp. 335-343.

Claverie, J. and Notredame, C. (2007), *Bioinformatics for dummies,* 2nd ed, Wiley, Hoboken, N.J.

Corpet, F. (1988), "Multiple sequence alignment with hierarchical clustering.", *Nucleic Acids Research,* vol. 16, no. 22, pp. 10881-10890.

Cserepes, J., Szentpétery, Z., Seres, L., O□zvegy-Laczka, C., Langmann, T., Schmitz, G., Glavinas, H., Klein, I., Homolya, L., Váradi, A., Sarkadi, B. and Elkind, N. B. (2004), "Functional expression and characterization of the human ABCG1 and ABCG4 proteins: Indications for heterodimerization", *Biochemical and Biophysical Research Communications,* vol. 320, no. 3, pp. 860-867.

Davidson, A. L. and Maloney, P. C. (2007), "ABC transporters: how small machines do a big job", *Trends in Microbiology,* vol. 15, no. 10, pp. 448-455.

Dayhoff,M.O., Schwartz, R.M., Orcutt, B.C. (1978), "A model for evolutionary change in proteins", *Atlas of Protein Sequence and Structure,* vol. 5, pp. 345-352.

Dean, M. and Annilo, T., (2005), *Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates*.

Dean, M., Hamon, Y. and Chimini, G. (2001a), "The human ATP-binding cassette (ABC) transporter superfamily", *Journal of Lipid Research,* vol. 42, no. 7, pp. 1007-1017.

Dean, M., Rzhetsky, A. and Allikmets, R. (2001b), "The human ATP-binding cassette (ABC) transporter superfamily", *Genome Research,* vol. 11, no. 7, pp. 1156-1166.

Edgar, R. C. (2004a), "MUSCLE: A multiple sequence alignment method with reduced time and space complexity", *BMC Bioinformatics,* vol. 5.

Edgar, R. C. (2004b), "MUSCLE: Multiple sequence alignment with improved accuracy and speed", pp. 728.

Efron, B. (1979),"Bootstrapping methods: Another look at the jacknife", *Annals of Statistics*, vol.7, pp.1-26.

Ensembl (2008), *Ensembl*, available at: http://www.ensembl.org/ (accessed 20th August 2008).

Felsenstein, J. (2005), "Using the quantitative genetic threshold model for inferences between and within species", *Philosophical Transactions of the Royal Society B: Biological Sciences,* vol. 360, no. 1459, pp. 1427-1434.

Felsenstein, J. (1985), "Confidence intervals on phylogenies: An approach using the bootstrap", *Evolution* vol.39, pp.783-791.

Gartner, J., Moser, H. and Valle, D. (1992), "Mutations in the 70K peroxisomal membrane protein gene in Zellweger syndrome.", *Nature genetics,* vol. 1, no. 1, pp. 16-23.

Gilbert, D. G. (2001), *Readseq2*, available at: http://iubio.bio.indiana.edu/soft/molbio/readseq/java (accessed 20th of August 2008).

GlaxoSmithkline Pharmaceuticals (2008), *Page GlaxoSmithKline Pharmaceuticals*, available at: http://www.gsk.com/ (accessed 20th of August 2008).

Graf, G. A., Yu, L., Li, W. -., Gerard, R., Tuma, P. L., Cohen, J. C. and Hobbs, H. H. (2003), "ABCG5 and ABCG8 Are Obligate Heterodimers for Protein Trafficking and Biliary Cholesterol Excretion", *Journal of Biological Chemistry,* vol. 278, no. 48, pp. 48275-48282.

Gregory, S. G., Sekhon, M., Schein, J., Zhao, S., Osoegawa, K., Scott, C. E., Evans, R. S., Burridge, P. W., Cox, T. V., Fox, C. A., Hutton, R. D., Mulienger, I. R., Phillips, K. J., Smith, J., Stalker, J., Threadgold, G. J., Birney, E., Wylie, K., Chinwalia, A., Wallis, J., Hillier, L., Carter, J., Galge, T., Jaeger, S., Kremitzki, C., Layman, D., Maas, J., McGrane, R., Mead, K., Walker, R., Jones, S., Smith, M., Asano, J., Bosdet, I., Chan, S., Chittaranjan, S., Chiu, R., Fjeil, C., Fuhrmann, D., Girn, N., Gray, C., Guin, R., Hsiao, L., Krzywinski, M., Kutsche, R., Lee, S. S., Mathewson, C., McLeavy, C., Messervier, S., Ness, S., Pandoh, P., Prabhu, A. -., Saeedi, P., Smallus, D., Spence, L., Stott, J., Taylor, S., Terpstra, W., Tsai, M., Vardy, J., Wye, N., Yang, G., Shatsman, S., Ayodeji, B., Geer, K., Tsegaye, G., Shvartsbeyn, A., Gebregeorgis, E., Kroi, M., Russell, D., Overton, L., Malek, J. A., Holmes, M., Heaney, M., Shetty, J., Feldblyum, T., Nierman, W. C., Catanese, J. J., Hubbard, T., Waterston, R. H., Rogers, J., De Jong, P. J., Fraser, C. M., Marra, M., McPherson, J. D. and Bentley, D. R. (2002), "A physical map of the mouse genome", *Nature,* vol. 418, no. 6899, pp. 743-750.

Gonnet, G. H., Cohen, M. A. and Benner, S. A. (1992), "Exhaustive matching of the entire protein sequence database", *Science,* vol. 256, no. 5062, pp. 1443-1445.

Gotoh, O. (2007), *PRRN*, available at: http://www.genome.ist.i.kyoto-u.ac.jp/~aln_user/prrn/index.html (accessed 20th of August 2008).

Heimer, S., Langmann, T., Moehle, C., Mauerer, R., Dean, M., Beil, F. U., von Bergmann, K. and Schmitz, G. (2002), "Mutations in the human ATP-binding cassette transporters ABCG5 and ABCG8 in sitosterolemia.", *Human mutation,* vol. 20, no. 2, pp. 151.

Henikoff, S. and Henikoff, J. G. (1992), "Amino acid substitution matrices from protein blocks", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 89, no. 22, pp. 10915-10919.

Higgins, C. F. (2001), "ABC transporters: Physiology, structure and mechanism - An overview", *Research in Microbiology,* vol. 152, no. 3-4, pp. 205-210.

Hillis, D. M. (1997a), "Biology recapitulates phylogeny", *Science,* vol. 276, no. 5310, pp. 218-219.

Hillis, D. M. (1997b), "Phylogenetic analysis", *Current Biology,* vol. 7, no. 3, pp. R129-R131.

Huelsenbeck, J. P. and Ronquist, F. (2001), "MRBAYES: Bayesian inference of phylogenetic trees", *Bioinformatics,* vol. 17, no. 8, pp. 754-755.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. and Bollback, J. P. (2001), "Bayesian inference of phylogeny and its impact on evolutionary biology", *Science,* vol. 294, no. 5550, pp. 2310-2314.

Huxley-Jones, J., Apte, S. S., Robertson, D. L. and Boot-Handford, R. P. (2005), "The characterisation of six ADAMTS proteases in the basal chordate Ciona intestinalis provides new insights into the vertebrate ADAMTS family", *International Journal of Biochemistry and Cell Biology,* vol. 37, no. 9, pp. 1838-1845.

Igarashi, Y., Aoki, K. F., Mamitsuka, H., Kuma, K. -. and Kanehisa, M. (2004), "The evolutionary repertoires of the eukaryotic-type ABC transporters in terms of the phylogeny of ATP-binding domains in eukaryotes and prokaryotes", *Molecular Biology and Evolution,* vol. 21, no. 11, pp. 2149-2160.

Jamroziak, K. and Robak, T. (2008), "Do polymorphisms in ABC transporter genes influence risk of childhood acute lymphoblastic leukemia?", *Leukemia Research,* vol. 32, no. 8, pp. 1173-1175.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992), "The rapid generation of mutation data matrices from protein sequences", *Computer Applications in the Biosciences,* vol. 8, no. 3, pp. 275-282.

Kalabis, G. M., Kostaki, A., Andrews, M. H., Petropoulos, S., Gibb, W. and Matthews, S. G. (2005), "Multidrug resistance phosphoglycoprotein (ABCB1) in the mouse placenta: Fetal protection", *Biology of Reproduction,* vol. 73, no. 4, pp. 591-597.

Kimura, M. (1980), "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences", *Journal of Molecular Evolution,* vol. 16, no. 2, pp. 111-120.

Koehn, J., Fountoulakis, M. and Krapfenbauer, K. (2008), "Multiple drug resistance associated with function of ABC-transporters in diabetes mellitus: Molecular mechanism and clinical relevance", *Infectious Disorders - Drug Targets,* vol. 8, no. 2, pp. 109-118.

Kolaczkowski, B. and Thomton, J. W. (2004), "Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogenous", *Nature,* vol. 431, no. 7011, pp. 980-984.

Kuhner, M. K. and Felsenstein, J. (1994), "A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates", *Molecular Biology and Evolution,* vol. 11, no. 3, pp. 459-468.

Lassmann, T. and Sonnhammer, E. L. L. (2002), "Quality assessment of multiple alignment programs", *FEBS Letters,* vol. 529, no. 1, pp. 126-130.

Li, Q., Jiang, Q., Larusso, J., Klement, J. F., Sartorelli, A. C., Belinsky, M. G., Kruh, G. D. and Uitto, J. (2007), "Targeted ablation of Abcc1 or Abcc3 in Abcc6-/- mice does not modify the ectopic mineralization process", *Experimental Dermatology,* vol. 16, no. 10, pp. 853-859.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas III, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., DeJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C. -., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K. -., Parker, H. G., Pollinger, J. P., Searle, S. M. J., Sutter, N. B., Thomas, R., Webber, C. and Lander, E. S. (2005), "Genome sequence, comparative analysis and haplotype structure of the domestic dog", *Nature,* vol. 438, no. 7069, pp. 803-819.

Liu, L. X., Janvier, K., Berteaux-Lecellier, V., Cartier, N., Benarous, R. and Aubourg, P. (1999), "Homo- and heterodimerization of peroxisomal ATP-binding cassette half- transporters", *Journal of Biological Chemistry,* vol. 274, no. 46, pp. 32738-32743.

Mackey, A.J., (2007), *VAST: Visualization of Aligned Sequences and Trees*.

Mickley, L., Jain, P., Miyake, K., Schriml, L. M., Rao, K., Fojo, T., Bates, S. and Dean, M. (2001), "An ATP-binding cassette gene (ABCG3) closely related to the multidrug transporter ABCG2 (MXR/ABCP) has an unusual ATP-binding domain", *Mammalian Genome,* vol. 12, no. 1, pp. 86-88.

Minoretti, P., Falcone, C., Aldeghi, A., Olivieri, V., Mori, F., Emanuele, E., Calcagnino, M. and Geroldi, D. (2006), "A novel Val734Ile variant in the ABCC9 gene associated with myocardial infarction", *Clinica Chimica Acta,* vol. 370, no. 1-2, pp. 124-128.

Mizutani, T., Masuda, M., Nakai, E., Furumiya, K., Togawa, H., Nakamura, Y., Kawai, Y., Nakahira, K., Shinkai, S. and Takahashi, K. (2008), "Genuine functions of P-glycoprotein (ABCB1)", *Current Drug Metabolism,* vol. 9, no. 2, pp. 167-174.

Morita, M. (2007), "Adrenoleukodystrophy: Molecular pathogenesis and development of therapeutic agents", *Yakugaku Zasshi,* vol. 127, no. 7, pp. 1059-1064.

Mount, D. W. (2004), *Bioinformatics : sequence and genome analysis,* 2nd ed, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Mourez, M., Jéhanno, M., Hofnung, M. and Dassa, E. (2000), "Role, functional mechanism and structure of ABC (ATP-binding cassette) transporters", *Medecine/Sciences,* vol. 16, no. 3, pp. 386-394.

National Center for Biotechnology Information and U.S. National Library of Medicine (2008), *National Center for Biotechnology Information (NCBI)*, available at: http://www.ncbi.nlm.nih.gov/ (accessed 20th of August 2008).

Needleman, S. B. and Wunsch, C. D. (1970), "A general method applicable to the search for similarities in the amino acid sequence of two proteins.", *Journal of Molecular Biology,* vol. 48, no. 3, pp. 443-453.

Nicholas, K. B., Nicholas, H. B. J. and Deerfield, D. W. (2007) *, GeneDoc: Analysis and Visualization of Genetic Variation*, available at: http://www.nrbsc.org/gfx/genedoc/index.html (accessed 20th August 2008).

Notredame, C. and Higgins, D. G. (1996), "SAGA: Sequence alignment by genetic algorithm", *Nucleic Acids Research,* vol. 24, no. 8, pp. 1515-1524.

Notredame, C., Higgins, D. G. and Heringa, J. (2000), "T-coffee: A novel method for fast and accurate multiple sequence alignment", *Journal of Molecular Biology,* vol. 302, no. 1, pp. 205-217.

Pahnke, J., Wolkenhauer, O., Krohn, M. and Walker, L. C. (2008), "Clinico-pathologic function of cerebral ABC transporters - Implications for the pathogenesis of Alzheimer's disease", *Current Alzheimer Research,* vol. 5, no. 4, pp. 396-405.

Perrière, G. and Gouy, M. (1996), "WWW-Query: An on-line retrieval system for biological sequence banks", *Biochimie,* vol. 78, no. 5, pp. 364-369.

Piehler, A. P., Wenzel, J. J., Olstad, O. K., Haug, K. B. F., Kierulf, P. and Kaminski, W. E. (2006), "The human ortholog of the rodent testis-specific ABC transporter Abca17 is a ubiquitously expressed pseudogene (ABCA17P) and shares a common 5′ end with ABCA3", *BMC Molecular Biology,* vol. 7.

Plomp, A. S., Florijn, R. J., ten Brink, J., Castle, B., Kingston, H., Martín-Santiago, A., Gorgels, T. G. M. F., de Jong, P. T. V. M. and Bergen, A. A. B. (2008), "ABCC6 mutations in pseudoxanthoma elasticum: An update including eight novel ones", *Molecular Vision,* vol. 14, pp. 118-124.

Rat Genome Sequencing Project Consortium (2004),"Genome sequence of the Brown Norway rat yields insights into mammalian evolution", *Nature,* vol. 428, no. 6982, pp. 493-520.

Ronquist, F. and Huelsenbeck, J. P. (2003), "MrBayes 3: Bayesian phylogenetic inference under mixed models", *Bioinformatics,* vol. 19, no. 12, pp. 1572-1574.

Saitou, N. and Nei, M. (1987), "The neighbor-joining method: a new method for reconstructing phylogenetic trees.", *Molecular biology and evolution,* vol. 4, no. 4, pp. 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", *Journal of Molecular Evolution,* vol. 48, no. 1, pp. 22-41.

Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V. and Altschul, S. F. (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", *Nucleic Acids Research,* vol. 29, no. 14, pp. 2994-3005.

Schmidt, H. A., Strimmer, K., Vingron, M. and Von Haeseler, A. (2002), "TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing", *Bioinformatics,* vol. 18, no. 3, pp. 502-504.

Smith, T. F. and Waterman, M. S. (1981), "Identification of common molecular subsequences", *Journal of Molecular Biology,* vol. 147, no. 1, pp. 195-197.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. and Birney, E. (2002), "The Bioperl toolkit: Perl modules for the life sciences", *Genome Research,* vol. 12, no. 10, pp. 1611-1618.

Struk, B., Neldner, K. H., Rao, V. S., St Jean, P. and Lindpaintner, K. (1997), "Mapping of both autosomal recessive and dominant variants of pseudoxanthoma elasticum to chromosome 16p13.1", *Human Molecular Genetics,* vol. 6, no. 11, pp. 1823-1828.

Subramanian, A. R., Weyer-Menkhoff, J., Kaufmann, M. and Morgenstern, B. (2005), "DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment", *BMC Bioinformatics,* vol. 6.

Szakács, G., Váradi, A., O□zvegy-Laczka, C. and Sarkadi, B. (2008), "The role of ABC transporters in drug absorption, distribution, metabolism, excretion and toxicity (ADME-Tox)", *Drug Discovery Today,* vol. 13, no. 9-10, pp. 379-393.

Taketani, S., Kakimoto, K., Ueta, H., Masaki, R. and Furukawa, T. (2003), "Involvement of ABC7 in the biosynthesis of heme in erythroid cells: Interaction of ABC7 with ferrochelatase", *Blood,* vol. 101, no. 8, pp. 3274-3280.

Tam, R. and Saier Jr., M. H. (1993), "Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria", *Microbiological Reviews,* vol. 57, no. 2, pp. 320-346.

The Chimpanzee Sequencing and Analysis Consortium (2005),"Initial sequence of the chimpanzee genome and comparison with the human genome", *Nature,* vol. 437, no. 7055, pp. 69-87.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997), "The CLUSTAL X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools", *Nucleic Acids Research,* vol. 25, no. 24, pp. 4876-4882.

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 Research,* vol. 22, no. 22, pp. 4673-4680.

Toyoda, Y., Hagiya, Y., Adachi, T., Hoshijima, K., Kuo, M. T. and Ishikawa, T. (2008), "MRP class of human ATP binding cassette (ABC) transporters: Historical background and new research directions", *Xenobiotica,* vol. 38, no. 7-8, pp. 833-862.

Velamakanni, S., Wei, S. L., Janvilisri, T. and Van Veen, H. W. (2007), "ABCG transporters: Structure, substrate specificities and physiological roles - A brief overview", *Journal of Bioenergetics and Biomembranes,* vol. 39, no. 5-6, pp. 465-471.

Whelan, S. (2008), "Inferring trees.", *Methods in molecular biology (Clifton, N.J.),* vol. 452, pp. 287-309.

Wilgenbusch, J. C. and Swofford, D. (2003), "Inferring evolutionary trees with PAUP*.", *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.],* vol. Chapter 6.

Xiong, J. (2006), *Essential bioinformatics,* Cambridge University Press, New York.

Yang, Z. (2007), "PAML 4: Phylogenetic analysis by maximum likelihood", *Molecular Biology and Evolution,* vol. 24, no. 8, pp. 1586-1591.

Yang, Z. (1997), "PAML: A program package for phylogenetic analysis by maximum likelihood", *Computer Applications in the Biosciences,* vol. 13, no. 5, pp. 555-556.

Appendices

5.2 Appendix I – List of ABC transporters gene references

Table 12. ABCA subfamily references used in the ABC transporters superfamily phylogenetic analysis (Chapter 2), for the species *Homo sapiens*, *Pan troglogytes*, *Rattus norvegicus*, *Mus musculus*, *Canis familiaris* and *Gallus gallus* (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008).

Gene	*Homo sapiens*	*Pan troglodytes*	*Rattus norvegicus*	*Mus musculus*	*Canis familiaris*	*Gallus gallus*
ABCA1	gbBC146856.1	XM_001138040.1	NM_178095.2	NM_013454.3	XM_538773.2	NM_204145.1
ABCA2	BC064542.1	XM_001168583.1	NM_024396.1	NM_007379.2	XM_537788.2	XM_422330.2
ABCA3	U78735.1HSU78735	XM_510744.2	XM_220219.4	NM_013855.2	XM_537004.2	XM_414701.2
ABCA4	gbAF001945.1AF001945	XM_001152577.1	NM_001107721.1	NM_007378.1	gbAY427779.1	XM_422330.2
ABCA5	dbjAK292592.1	XM_001166579.1	embAJ550165.1RNO550165	NM_147219.2	XM_857705.1	XM_415695.2
ABCA6	gbBC125231.1	XM_001146278.1	XM_001081607.1	gbBC132417.1	XM_845829.1	
ABCA7	dbjAB055390.1	XM_524026.2	NM_207598.1	gbBC024511.1	XM_542208.2	
ABCA8	BC130280.1	XM_001166010.1	XM_001081601.1	NM_013851.1	XM_548020.2	XM_415691.2
ABCA9	NM_080283.3	XM_001146190.1	XM_001081605.1	gbAF491299.1	XM_848625.1	
ABCA10	NM_080282.3	XM_001165871.1				
ABCA11	dbjAK024359.1					XM_415691.2
ABCA12	NM_015657.3	XM_001149590.1	XM_237242.4	XM_987225.1	XM_857905.1	XM_421867.2
ABCA13	gbAY204751.1	XM_001147042.1	NM_001106020.2	NM_178259.3	XM_843318.1	XR_027192.1
ABCA14			XM_001079182.1	gbAY243470.1		
ABCA15			NM_001106293.1	gbBC141350.1		
ABCA17	NR_003574.1	XM_523266.2	NM_001031637.1			

Table 13. ABCB subfamily references used in the ABC transporters superfamily phylogenetic analysis (Chapter 2), for the species Homo *sapiens*, *Pan troglogytes*, *Rattus norvegicus*, *Mus musculus*, *Canis familiaris* and *Gallus gallus* (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008).

Gene	*Homo sapiens*	*Pan troglodytes*	*Rattus norvegicus*	*Mus musculus*	*Canis familiaris*	*Gallus gallus*
ABCB1	dbjAK290159.1	XM_519183.2	NM_133401.1	NM_011076.2	gbDQ068953.1	NM_204894.1
TAP1 / ABCB2	gbEU176425.1	XM_001166781.1	gbBC101854.1	NM_013683.1	embAJ630364.1	
TAP2 / ABCB3	dbjAK222823.1	XR_022058.1	NM_032056.2	gbU60087.1MMU60087	XM_532099.2	
ABCB4	NM_018849.2	XM_001160982.1	NM_012690.1	NM_008830.2		XM_418636.2
ABCB5	gbAY230001.1	XM_001152831.1	XM_001062082.1	dbjAK020318.1	XM_539461.2	
ABCB6	gbDQ895063.2	XM_001161097.1	gbBC085712.1	dbjAK168642.1	XM_536073.2	XM_423449.2
ABCB7	gbBT009918.1		NM_212518.1	dbjAK151967.1	XM_549087.2	XM_420301.2
ABCB8	gbAF047690.1AF047690	XM_519524.2	gbBC085781.1	NM_029020.2	XM_539916.2	
ABCB9	gbBC017348.2	XM_509453.2	dbjAB116265.1	NM_019875.2	XM_853575.1	XM_415125.2
ABCB10	gbAF216833.1AF216833	XR_023717.1	NM_001012166.1	NM_019552.2		XM_419578.2
ABCB11	NM_003742.2	XM_526100.2	NM_031760.1	gbAF133903.1AF133903	XM_545512.2	XR_027131.1

Table 14. ABCC subfamily references used in the ABC transporters superfamily phylogenetic analysis (Chapter 2), for the species *Homo sapiens*, *Pan troglogytes*, *Rattus norvegicus*, *Mus musculus*, *Canis familiaris* and *Gallus gallus* (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008).

Gene	*Homo sapiens*	*Pan troglodytes*	*Rattus norvegicus*	*Mus musculus*	*Canis familiaris*	*Gallus gallus*
ABCC1	gbBC157105.1	XM_001145351.1	gbAY170916.1	NM_008576.2	NM_001002971.1	NM_001012522.1
ABCC2	NM_000392.3	XM_507976.2	NM_012833.1	NM_013806.2	gbAY582532.1	XM_421698.2
ABCC3	gbAF085692.1AF085692	XM_001158914.1	NM_080581.1	NM_029600.3	XM_548204.2	XM_420102.2
ABCC4	AY081219.1	XM_001136373.1	NM_133411.1	NM_001033336.2	XM_542642.2	NM_001030819.1
ABCC5	gbBC050744.1	XR_022843.1	gbBC128730.1	NM_176839.1	XM_852305.1	XM_422754.2
ABCC6	NP_001162.4	XM_001166102.1	dbjAB010466.1	dbjAB028737.1	XM_547113.2	XM_001234743.1
CFTR/ ABCC7	NM_000492.3	XM_519330.2	XM_519330.2	NM_031506.1	gbM69298.1MUSCFTR	NM_001007143.1
ABCC8	gbU63421.1HSU63421	XM_508310.2	dbjAB052294.1	gbBC141411.1	XM_542520.2	XM_001232386.1
ABCC9	NM_020298.2	XM_001149494.1	gbAF087838.1AF087838	NM_001044720.1	XM_543765.2	XR_027154.1
ABCC10	gbBC166699.1	XM_518494.2	NM_001108201.1	NM_145140.2	XM_538934.2	XM_419506.2
ABCC11	gbBC039085.1	XM_001163474.1			XM_858602.1	
ABCC12	gbAF411578.1	XM_001163361.1	NM_199377.1	gbBC138381.1	XM_544420.2	
ABCC13	gbAY063514.1					XM_416677.2

Table 15. ABCD, ABCD, ABCE and ABCF subfamilies references used in the ABC transporters superfamily phylogenetic analysis (Chapter 2), for the species *Homo sapiens*, *Pan troglogytes*, *Rattus norvegicus*, *Mus musculus*, *Canis familiaris* and *Gallus gallus* (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008).

Gene	*Homo sapiens*	*Pan troglodytes*	*Rattus norvegicus*	*Mus musculus*	*Canis familiaris*	*Gallus gallus*
ABCD1	gbDQ894177.2	XR_023007.1	NM_001108821.1	gbBC079840.1	XM_850248.1	
ABCD2	NM_005164.3	XM_001168647.1	NM_033352.1	dbjAK082588.1	XM_859591.1	XM_415938.2
ABCD3	gbM81182.1HUMPMP	XM_513575.2	NM_012804.1	gbBC054446.1	XM_537064.2	NM_001012597.1
ABCD4	dbjAK291332.1	XM_510061.2	XM_001059055.1	NM_008992.1	XM_547903.2	XM_421264.2
ABCE1	gbBT009779.1	XM_517465.2	NM_001108446.1	NM_015751.2	XM_532679.2	NM_001006440.1
ABCF1	NM_001090.2	NM_001042379.1	gbBC100256.1	gbBC046965.1	XM_532056.2	
ABCF2	embCU674820.1	XM_001139681.1	gbBC129119.1	dbjAK087990.1	XM_855899.1	NM_001006562.1
ABCF3	NM_018358.2	XM_516910.2	NM_001011896.1	NM_013852.2	XM_853965.1	XM_422757.2
ABCG1	NM_004915.3	XM_514918.2	NM_053502.1	gbBC119471.2	XM_544902.2	embBX934614.1
ABCG2	gbDQ895507.2	XM_526633.2	gbAY089998.1	gbBC053730.1	NM_001048021.1	XM_421638.2
ABCG3			NM_001004076.2	dbjAK156626.1		
ABCG3s			NM_001037205.1			
ABCG4	embAJ308237.1HSA308237	XM_522202.2	NM_001109883.1	NM_138955.3	XM_848138.1	embAJ308237.1HSA308237
ABCG5	NM_022436.2	XR_023446.1	NM_053754.2	gbAY195873.1	XM_538475.2	XM_419457.2
ABCG8	gbAF320294.1AF320294	XM_525745.2	NM_130414.2	gbBC138484.1	XM_531799.2	XM_419458.2

5.3 Appendix II - Human ABC transporters – Genetic Location, Function and Associated diseases.

Table 16. ABCA subfamily. Official Symbol Gene id, Aliases, Location, Function and Associated diseases functional and disease associated information (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008) (June 2008). Functions of genes unknown are left in blank.

Official Symbol	Location	Specific Function	Associated diseases
ABCA1	9q31.1	Cholesteral efflux pump in the cellular lipid removal pathway	Tangier's disease and familial high-density lipoprotein deficiency
ABCA2	9q34	Highly expressed in brain tissue and may play a role in macrophage lipid metabolism and neural development	
ABCA3	16p13.3	May be involved in development of resistance to xenobiotics and engulfment during programmed cell death	
ABCA4	1p22.1-p21	Retina-specific ABC transporter with N-retinylidene-PE as a substrate.	Stargardt disease, retinitis pigmentosa-19, cone-rod dystrophy type 3, early-onset severe retinal dystrophy, fundus flavimaculatus, and macular degeneration age-related 2
ABCA5	17q24.3		
ABCA6	17q24.3	May play a role in macrophage lipid homeostasis	
ABCA7	19p13.3	May play a role in lipid homostasis in cells of the immune system	
ABCA8	17q24		
ABCA9	17q24.2		
ABCA10	17q24		
ABCA11	4p16.3	Pseudogene	
ABCA12	2q34		
ABCA13	7p12.3		
ABCA17	16p13.3	Pseudogene	

Table 17. ABCB subfamily. Official Symbol Gene id, Aliases, Location, Function and Associated diseases functional and disease associated information (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008) (June 2008).

Official Symbol	Location	Specific Function	Associated diseases
ABCB1	7q21.1	ATP-dependent drug efflux pump for xenobiotic compounds with broad substrate specificity. It is responsible for decreased drug accumulation in multidrug-resistant cells and often mediates the development of resistance to anticancer drugs	Renal function ; nephropathy
TAP1 / ABCB2	6p21.3	Pumping of degraded cytosolic peptides across the endoplasmic reticulum into the membrane-bound compartment where class I molecules assemble	Ankylosing spondylitis, insulin-dependent diabetes mellitus, and celiac disease
TAP2 / ABCB3	6p21.3	Peptide transport from the cytoplasm to the endoplasmic reticulum	Ankylosing spondylitis, insulin-dependent diabetes mellitus, and celiac disease
ABCB4	7q21.1	May play a role in the transport of phospholipids from liver hepatocytes into bile	
ABCB5	7p15.3	May play a role in lysosomes	
ABCB6	2q36	May play a role in mitochondrial function	Candidate gene for lethal neonatal metabolic syndrome, a disorder of mitochondrial function
ABCB7	Xq12-q13	Transport of heme from the mitochondria to the cytosol. This protein may play a role in metal homeostasis	X-linked sideroblastic anemia with ataxia
ABCB8	7q36	May play a role in the compartmetalisation and transport of heme,as well as peptides, from the mitochondria to the nucleus and cytosol	
ABCB9	12q24	May play a role in lysosomes	
ABCB10	1q42.13		
ABCB11	2q24	The major canalicular bile salt export pump in man	Progressive familial intrahepatic cholestases which are a group of inherited disorders with severe cholestatic liver disease from early infancy

Table 18. ABCC subfamily. Official Symbol Gene id, Aliases, Location, Function and Associated diseases functional and disease associated information (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008) (June 2008).

Official Symbol	Location	Specific Function	Associated diseases
ABCC1	16p13.1	Multispecific organic anion transporter	
ABCC2	10q24	Biliary transport and drug resistance in mammalian cells	Dubin-Johnson syndrome (DJS)
ABCC3	17q22	May play role in the transport of biliary and intestinal excretion of organic anions	
ABCC4	13q32	May play a role in cellular detoxification	
ABCC5	3q27	Contributes to the degradation of phosphodiesterases and may play a role int the elimination pathway for cyclic nucleotides	Involved in resistance to thiopurines in acute lymphoblastic leukemia and antiretroviral nucleoside analogs in HIV-infected patients
ABCC6	16p13.1		Pseudoxanthoma elasticum
CFTR/ ABCC7	7q31.2	Chloride channel and controls the regulation of other transport pathways	Autosomal recessive disorders cystic fibrosis and congenital bilateral aplasia of the vas deferens
ABCC8	11p15.1	Modulator of ATP-sensitive potassium channels and insulin release	Hyperinsulinemic hypoglycemia of infancy and non-insulin-dependent diabetes mellitus type II
ABCC9	12p12.1	May play a role as the drug-binding channel-modulating subunit of the extrapancreatic ATP-sensitive potassium channels	
ABCC10	6p21		
ABCC11	16q12.1	Plays a role in physiological processes with bile acids, conjugated steroids, and cyclic nucleotides	Earwax type determination
ABCC12	16q12.1		Increased expression of this gene is associated with breast cancer.
ABCC13	21q11.2		

Table 19. ABCD, ABCF, ABCG subfamilies. Official Symbol Gene id, Aliases, Location, Function and Associated diseases functional and disease associated information (National Center for Biotechnology Information and U.S. National Library of Medicine, 2008) (June 2008).

Official Symbol	Location	Specific Function	Associated diseases
ABCD1	Xq28	May play a role in peroxisomal transport or catabolism of very long chain fatty acids	Adrenoleukodystrophy
ABCD2	12q11-12		Adrenoleukodystrophy, Zellweger syndrome
ABCD3	1p22-p21	Role in peroxisome biogenesis	Zellweger syndrome
ABCD4	14q24.3	May play a role in adrenoleukodystrophy phenotype modification	Adrenoleukodystrophy
ABCE1	4q31	Blocks the activity of ribonuclease L	
ABCF1	6p21.33	May play a role in enhancement of protein synthesis and the inflammatio process.	
ABCF2	7q36		
ABCF3	3q27.1		
ABCG1	21q22.3	Macrophage cholesterol and phospholipids transport	
ABCG2	4q22	Xenobiotic transporter which may play a role in multi-drug resistance	
ABCG4	11q23.3	May play a role in cholesterol transport	
ABCG5	2p21	To limit intestinal absorption and promote biliary excretion of sterols	Sterol accumulation , atheroschlerosis and sitosterolemia
ABCG8	2p21	To exclude non-cholesterol sterol entry at the intestinal level, promote excretion of cholesterol and sterols into bile, and to facilitate transport of sterols back into the intestinal lumen	Sterol accumulation , atheroschlerosis and sitosterolemia