A Computational Approach From Gene to Structure Analysis of the Human ABCA4 Transporter Involved in Genetic Retinal Diseases

Alfonso Trezza,1 Andrea Bernini,1 Andrea Langella,1 David B. Ascher,2 Douglas E. V. Pires,3 Andrea Sodi,4 Ilaria Passerini,5 Elisabetta Pelo,5 Stanislao Rizzo,4 Neri Niccolai,1 and Ottavia Spiga1

1Department of Biotechnology Chemistry and Pharmacy, University of Siena, Siena, Italy
2Department of Biochemistry and Molecular Biology, University of Melbourne, Bio21 Institute, Parkville, Australia
3Instituto René Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Brazil
4Department of Surgery and Translational Medicine, Eye Clinic, Careggi Teaching Hospital, Florence, Italy
5Department of Laboratory Diagnosis, Genetic Diagnosis Service, Careggi Teaching Hospital, Florence, Italy


PURPOSE. The aim of this article is to report the investigation of the structural features of ABCA4, a protein associated with a genetic retinal disease. A new database collecting knowledge of ABCA4 structure may facilitate predictions about the possible functional consequences of gene mutations observed in clinical practice.

METHODS. In order to correlate structural and functional effects of the observed mutations, the structure of mouse P-glycoprotein was used as a template for homology modeling. The obtained structural information and genetic data are the basis of our relational database (ABCA4Database).

RESULTS. Sequence variability among all ABCA4-deposited entries was calculated and reported as Shannon entropy score at the residue level. The three-dimensional model of ABCA4 structure was used to locate the spatial distribution of the observed variable regions. Our predictions from structural in silico tools were able to accurately link the functional effects of mutations to phenotype. The development of the ABCA4Database gathers all the available genetic and structural information, yielding a global view of the molecular basis of some retinal diseases.

CONCLUSIONS. ABCA4 modeled structure provides a molecular basis on which to analyze protein sequence mutations related to genetic retinal disease in order to predict the risk of retinal disease across all possible ABCA4 mutations. Additionally, our ABCA4 predicted structure is a good starting point for the creation of a new data analysis model, appropriate for precision medicine, in order to develop a deeper knowledge network of the disease and to improve the management of patients.

Keywords: ABCA4 protein, ABC transporter, mutations database

ABCA4, also known as ABCR or the Rim protein, is a retinal-specific member of the ABC subfamily of adenosine triphosphate (ATP)-binding cassette (ABC) transporters. ABC transporters constitute one of the largest protein superfamilies. These evolutionarily, highly conserved transmembrane proteins use the energy of ATP hydrolysis to translocate a broad cluster of molecules across cell membrane, functioning as either importers or exporters. ABCA4 is predominantly expressed in vertebrate rod and cone photoreceptor cells, and it is mainly localized in outer segment disk membranes; the protein is also marginally present in the brain.

In recent years, significant progress has been made in our understanding of roles and mechanisms for ABCA4 activity. Several biochemical studies have been carried out to determine its natural substrate, now accepted to be all-trans-retinal and N-retinylidene-phosphatidylethanolamine (N-retinylidene-PE), a conjugate of all-trans-retinal and phosphatidylethanolamine.

Photoreceptor-specific expression of ABCA4 and its localization confirms the suggestion that ABCA4 may transport a substrate that is critical for photoreceptor function. Analysis of ABCA4 knockout mice together with biochemical studies, has led to a conceptual model for the role of ABCA4 as a retinoid transporter in the visual cycle. This suggests that ABCA4 may act as an ATP-dependent flippase that translocates N-retinylidene-PE, promoting the recycling of all-trans-retinal released from photobleached rhodopsin via the retinoid cycle. Compromised ABCA4 transporter activity may lead to the accumulation of toxic all-trans-retinal derivatives in rods and cones, leading to apoptosis of the supporting retinal pigment epithelium cells and, ultimately, degeneration of photoreceptors.

ABCA4 was the first ABCA-transporter that has been linked to various genetic retinal diseases, such as (Stargardt disease, cone dystrophy, and retinitis pigmentosa). Some ABCA4 sequence variants also have been considered as a risk factor for the development of AMD, a frequent progressive degener-
ation of central retina, with a high prevalence in elderly people.26,27

Stargardt disease (STGD1) is the most common inherited juvenile macular degeneration and is associated with a reduction of central visual acuity.28 On funduscopy, it is characterized by macular atrophy, often surrounded by typical fishtail yellowish deposits at the posterior pole and in midperiphery (Reckts).29 Cone dystrophy is characterized by the degeneration of cone photoreceptors with photophobia, reduced visual acuity, and abnormal color vision.30,31 Fundus appearance may sometimes be normal, but it may often show a macular dystrophy. The main feature of retinitis pigmentosa is progressive photoreceptor atrophy (first rods and then cones), with night blindness and visual acuity and visual field loss. In most cases, patients show typical pigmented deposits on fundus examination.32,33 The association of ABCA4 with AMD is more controversial, as some studies report higher frequency of ABCA4 mutations for AMD in comparison with a control population,24,25 even though these results have not been confirmed by other investigations.34 Human ABCA4 is a large single polypeptide of 2273 amino acids organized as two structurally related tandem-arranged halves.1 Each half consists of three structural domains, where the first is a transmembrane domain (TMD) comprising six transmembrane α-helices (TMs), two single TMs are at positions 22–42 (TM1) and 1377–1397 (TM7), delineating the beginning of the two halves, respectively. The second hydrophobic region is formed by the other five TMs, that is, TM2-6 at positions 653–899 and TM8-12 at positions 1670–1964. The second domain is a lumen domain, containing a large exocytoplasmic domain (ECD), that is, ECD1 50–645 and ECD2 1442–1644 positioned between TM1 and TM2-6 and TM7 and TM8-12.35 The third domain, located on the cytoplasmic side of disk membrane at positions 958–1176 and 1934–2147, respectively,36 is a large soluble-protein region containing the nucleotide-binding domain (NBD), responsible for ATP hydrolysis. TMDs are homologous only with a small subfamily of ABC transporters, containing the drug-binding sites and forming the structural system for translocation.12,35,36 ECDs do not show a significant sequence similarity to known proteins, except for other ABCA4 sequences from different species. They are glycosylated37 and are probably involved in the interactions with other proteins.38 NBDs are the only regions of ABC transporters that are structurally highly conserved.12 Based on functional and mutagenesis studies, NBD1 is responsible for basal ATPase activity, whereas NBD2 produces the retinal-stimulated increase in activity.38,39 Suggested mechanism of transport proposed is based on the alternating access model, established for smaller ABC trans-

MATERIALS AND METHODS

Reconstruction of the Phylogenetic Tree and Residue Conservation

ABCA4 gene and amino acid sequence analyses were performed by using the following software packages and Web-accessible programs: NCBI Blast,52 MUSCLE,53 COBALT,54 and Scorecons.55,56 In order to reconstruct a phylogenetic tree of ABCA4, we searched for homologs using PSI-BLAST.52 The resulting 100 sequences were aligned with MUSCLE,53 and this alignment was fed to COBALT54 to build a complete phylogenetic tree. The evaluation of residue conservation among ABCA4 sequences was carried out on the basis of a subset of nine whole protein sequences aligned with MUSCLE.53 Residue conservation was scored by using the Scorecons55 server. Sequence variability, $s_p$, where $s_p = 1 – Ss$ and Ss is the normalized Shannon's sequence entropy score given by Scorecons, was calculated for the entire ABCA4 human sequence. The scoring method is based on a classification of residues into one of seven types, a convention that follows that of Mirny and Shakhnovich.7

Model Building and Validation

Chain A from the crystal structure of mouse Pgp63 (PDB ID: 3G60) was used as the main template for the homology model. The UniProt database (an open-source database by the UniPort Consortium, http://www.uniprot.org, in the public domain) sequences of mouse Pgp (P06795) and human ABCA4 (P78365) were used for alignment. The alignment was obtained with the Clustal X tool58 using the default parameters with the blosum62 substitution matrix.59 The percentage of similarity (45%) between the two sequences makes this template a potential candidate to elaborate a 3D model of ABCA4 by comparative modeling. Each ABCA4 half-transporter was modeled separately using DeepView/Swiss-PdbViewer v. 3.7 (SPDV).60 Additional loop regions of ABCA4 were...
reconstructed from the Loops in Proteins database. The subunit linker was not modeled. Instead, ECDs were modeled using the threading bioinformatic approach with Phyre v.2.61 To each ABCA4, half the corresponding ECD subunit was linked and then the half-transporter models were reconstructed as a dimer to reproduce the ABCA4 subunit interface, refined to remove steric clashes upon energy minimization. All the images were prepared with PyMOL.62

The backbone root-mean-square deviation (RMSD) between Pgp (3G61)63 and the ABCA4 model was 2.5 nm. Due to the lack of a template structure, the RMSD could not be calculated for residues 50–645, 900–957, and 1442–1644. The quality of the modeled domains was assessed using local factors such as packing quality, backbone conformation, bond length, and side chain planarity. The Protein Model Portal64 showed that the ABCA4 model was of comparable quality to Pgp, while the analysis of the Ramachandran plot in Swiss-PdbViewer showed that over 90% of the residues fall within favorable regions, increasing our confidence in the backbone geometry. The obtained ABCA4 model structure was fitted by using Chimera software v.1.10.160 into the EMD ID 5497 with a reported resolution of 18 Å.55

Predicting Protein Structure Perturbation

Three computational methods were used to predict and analyze the effects of over 370 missense mutations on a structural basis: mCSM,66 Site-Directed Mutator,67 and DUET.68 The effect of the mutations was assessed in the context of the molecular interactions of the wild-type residues, and those programs were used to predict the effects of the mutations on protein thermal stability.

ABCA4Database

We manually curated a comprehensive table of germline and somatic ABCA4 mutations obtained from numerous sources, including original articles, the retinal disease related to ABCA4 from different public sources as Retina International Database,69 the Human Gene Mutation Database,70 and LOVD ABCA4 Database.71 Details of mutations not included in these sources were obtained from the original reference by Department of Laboratory Diagnosis, Genetic Diagnosis Service, Careggi Teaching Hospital. For each missense mutation, the new database shows different features: phenotype of retinal disease, sequence entropy/variability, structural localization, amino-acidic accessibility, structural stability, ligand binding, genetic correlations, references, etc. MySQL falls into the RDBMS family, a synonymous term for Relational Database Management System. In particular, it is entirely open source, and it has been the world’s most widely used RDBMS, open-source and client-server model RDBMS. SQL stands for Structured Query Language. MySQL is frequently chosen for use in web applications and is a central component of the widely used LAMP open-source web application software stack (LAMP, a synonym for Linux, Perl/PHP/Python).

Results

We first constructed, with the available experimental and phenotypic data, an inclusive database of ABCA4 missense mutations (ABCA4Database, http://www.sbl.unisi.it/abca4/abcr_mainlist.php, in the public domain). In order to identify the most important residues for ABCA4 structure and function, we have investigated protein residue conservation among a subset of homolog ABC proteins in several vertebrates. Understanding the most conserved regions of the protein would suggest the role of each mutation for the various retinal diseases. The first step was to search for ABCA4 homologs using PSI-BLAST.52 The 100 obtained sequences were selected to perform multiple alignments and to reconstruct a phylogenetic tree (Fig. 1) using the COBALT tool.54

It is apparent that the sequences that are closest to human ABCA4 are the nine ABCA4 proteins belonging to several primates. We decided to focus on these sequences and to analyze the residue conservation using Shannon’s information theoretical entropy.35,56 Accordingly, we have assessed the level of sequence conservation related to each position in the alignment, the similarity, identity, and entropy per position. Shannon entropy value (H) reflects the variability of every residue between compared sequences, and consequently, their conservation. The returned scores are normalized in order to have conserved columns with a score of 1 and diverse columns with a score of 0. For our analysis, these values, ranging from 0 to 1, were inverted, with 0 for invariant and 1 for hypervariable protein sequence positions, respectively. Thus, we have analyzed the entire ABCA4 sequence to gain information on the presence of critical positions along the polypeptide chain. The obtained entropy plot allowed us to have an idea of the amount of variability at each residue position within the alignment, having a measure of the lack of predictability for an alignment position. After that, sequence variability, s_j, where s_j = 1 – S_j and S_j is the normalized Shannon’s sequence entropy score, was calculated, and results for each ABCA4 domain are shown in Figure 2.

ECD1 and ECD2 contain, respectively, 145 (24%) and 105 (32%) residues having s_j = 0. These values increase to 51 (40%) and 63 (50%) for TMD1 and TMD2, respectively, and to 124 (53%) and 120 (51%) for NBD1 and NBD2, respectively. The values are up to 304 (49%) and 178 (54%) for ECD1 and ECD2, to 75 (62%) and 84 (70%) for TMD1 and TMD2, and to 158 (69%) and 172 (75%) for NBD1 and NBD2 for s_j smaller than 0.2. This finding shows that, although conserved residues exist in all six domains, it is evident that the TMDs and NBDs are the most highly conserved. We decided to analyze the distribution of sequence entropy by using the quartiles method, and in particular, we considered the value of the third quartile (Q3). Q3 values were obtained for the entire sequence distribution and for each individual domain, as reported in Figure 3.

The Q3 overall sequence value was 0.29, consistent with the singular regions of Figure 3, while for some parts it is much smaller, as in the case of TM2, TM5, and TM10 helices with 0.05, 0.02, and 0.01 Q3 values, respectively, and the two NBDs that have Q3 values of 0.17 and 0.13.

The analysis of sequence entropy within each domain allows us to understand which are the regions of ABCA4 most susceptible to the action of a single mutation, and considering the incidence of missense variants, we can obtain the maximum for the position with low-entropy values, that is, for highly conserved residues. Consequently, the incidence of mutations appears to have a downward trend in correspondence with increasing entropy values, demonstrating that most disease mutations were located at conserved residue positions.

To understand the spatial agreement between the sequence and the structure of a protein and to explore the influence of mutations associated with retinal disease on the structure and functionality of ABCA4 protein, a homology model was generated (see Materials and Methods). To obtain an accurate model of ABCA4 we separately considered the three domains of each half of the protein. The TMDs and NBDs were modeled using as a template the structure of mouse Pgp (PDB ID code 3G60 chain A), a member of the ABC superfamily.65 The structure of Pgp represents a nucleotide-free inward-facing conformation arranged as two “halves.” Since an important aspect of homology modeling is the sequence alignment
between the template and target proteins, we highlight the reasonable degree of sequence identity between the two proteins (35% for NBDs and 30% for TMDs). The mouse P-gp could not be used as a template for ECDs because suitable sequence alignment with ABCA4 could not be found. Thus, a different ABCA4 homology modeling approach for this region was adopted. A full atom model for NBDs and TMDs was generated by manual alignment between the secondary structure elements predicted by PsiPred runs and those observed in the template (3G60), enhancing the correspondence of identical and positively conserved residues of about 40% and 70% for NBDs and 35% and 60% for TMDs. On the other hand, the ECDs were modeled via the threading bioinformatic approach with Phyre v. 2.0. The geometrical and stereochemical features were calculated for validation by the Protein Model Portal validation system. The results show that more than 90% of the torsion angles were within the expected Ramachandran regions, all bond distances and angles

---

**Figure 1.** The phylogenetic tree was obtained from multiple alignment of 100 homology sequences of human ABCA4 by using as parameters: tree method COBALT Tree; Max Seq Difference 0.85; Distance Grishin (protein); Sequence Label Taxonomic Name (if available); Collapse Mode Blast Name. Arrows represent a branch of collapsed multisequences, and dots represent single sequences.

**Figure 2.** Histogram of residues for each structural domain. Blue bins represent the number of amino acids with $S_{co}$ variability equal to 0, yellow bins refer to the amino acid dimension of each domain. Arrows highlight the structural domains with a total number of conserved positions higher than 50%.
lie within the allowable range of the standard dictionary values, and the atom chirality of the models is correct, indicating that our ABCA4 model is reasonable in terms of geometry and stereochemistry.

The primary 3D model of ABCA4 (ECDs + TMDs + NBDs) from the structural modeling is shown in Figure 4.

Recently, the Electron Microscopy Map of ABCA4 in complex with adenyl-imidodiphosphate (AMPPNP) was published in the EMDB (http://www.emdatabank.org, in the public domain) with ID EMD-5498.72 The 3D structural analysis of ABCA4 by electron microscopy shows, quite clearly, the three portions of our protein: ECDs, TMDs, and NBDs.65 By using Chimera v.1.10.176 “Fit in Map” command, our 3D-modeled coordinates were transformed by fitting with the EMD-5498 density map into a plot whose 0.71 cross-correlation coefficient suggests a good reliability of our results (Fig. 5). The obtained superposition can be used to get an overall idea of the correct architecture of our model.

We then tried to predict possible pathogenic damages, using mCSM66 and DUET,68 with the structural analysis of missense mutations as collected in our ABCA4 Database, allowing their classification based on special deleterious effects due to backbone rigidity or, conversely, loss of backbone
rigidity, which can alter structural stability or residues' binding capability (Fig. 6).

**DISCUSSION**

We have initially focused our attention on testing residue conservation along the whole ABCA4 protein sequence. This is a crucial step in order to identify the most important regions for protein functionality and thus to understand the effects that mutations associated with retinal disease might have on it. Mutations occurring in a conserved site are assumed to have a greater impact on the correct protein folding and, consequently, to alter its functionality. Indeed, phylogenetically conserved sequences in divergent species probably identify a protein region of extreme importance for structural stability. On the contrary, in sequence regions exhibiting low-residue conservation, the effect of a mutation could not severely damage protein structure, possibly preserving its normal activity. By scoring sequence entropy for ABCA4, we observed
that in general the TMDs and the NBDs show about 20% more residue conservation than do the ECDs (Fig. 5). This is not unexpected, as TMDs are domains that keep the ABCA4 protein embedded in the membrane and NBDs are responsible for ATP binding, a key process for the retinal translocation. On the other hand, the ECDs do not exhibit the same degree of residue conservation and show only some consecutive residues with a very low sequence variability. This leads us to suggest that mutations in those domains are more likely to be tolerated. ECDs participating in substrate translocation process with their highly mobile hinge domains can more easily accommodate mutations without disrupting functionality. The 3D structure analysis of the ABCA4 protein is necessary to understand location and spatial distribution of the observed variable regions. In the absence of high-resolution experimentally derived structures, homology models provide a suitable 3D protein map; of course, as structural predictions, homology models require progressive refinement throughout whenever new structural/functional results become available. The crucial factors determining the quality of homology models are the sequence alignment accuracy and the quality and resolution of the template structure.

Currently, there are nine x-ray structures for full-length ABC transporters, but only two are of mammalian origin: the human ABCB1060 and the mouse ABCB1 81 also known as P-gp. We then decided to use as a template the mouse transporter ABCB1080 and the mouse ABCB1, 81 also known as P-gp. The Pgp (ABCB1) and ABCA4 are members of the same protein family, having the same architecture, except for the exocyttoplasmic regions that are absent in Pgp. For this reason, we have used the sequence alignment as a strong basis for accurate homology modeling of TMDs and NBDs; ECDs, instead, were modeled with the threading bioinformatic approach.61,73

Thus, the obtained 3D structure is a first approximation to the actual protein structure. Nevertheless, the predicted structure may constitute a useful starting point to understand the complete structural picture of ABCA4 at the different stages of the catalytic cycle, to characterize its functional states, to identify possible conformational changes, and to detect binding regions and cavity pockets. The latter features can guide further investigations on the role of residues that are located at the NBD:NBD, TMD:TMD, and NBD:TMD interfaces, explaining experimental data coming from mutation, cross-linking, and photo-labeling studies. Based on our 3D protein structure, we can differentiate between residues directly involved in particular functions and possibly influenced residues due to their close proximity. For every single mutation, the ABCA4Database contains information about the 3D local structure and the corresponding Shannon’s entropy,55 yielding results for the various ABCA4 domains with different reliabilities, which are low for ECDs. Nevertheless, the ABCA4 model offers a powerful template to predict the effects of variants responsible for ABCA4-related diseases. Thus, a better understanding of the correlation between a single mutation and a specific pathogenic phenotype, with significant prognostic implications in clinical practice, will be possible, including the identification of new drug-binding sites, supporting the efforts toward medical treatment of genetic retinal diseases.

Acknowledgments

The authors thank Francesco Poggialini and Simone Cirri for their valuable assistance in protein modeling procedures. The authors also thank Elena and Dario Sennati for their insightful contribution to this work.

Supported by the Jack Brockhoff Foundation (JBF 4186) (DBA), a C.J. Martin Research Fellowship from the National Health and Medical Research Council of Australia (APP1072476) (DBA), the René Rachou Institute (IRR/FIOCRUZ Minas) (DEVP), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (DEVP), and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (DEVP).

Disclosure: A. Trezza, None; A. Bermini, None; A. Langella, None; D.B. Ascher, None, D.E.V. Pires, None; A. Sodi, None; I. Passerini, None; E. Pelo, None; S. Rizzo, None; N. Nicolai, None; O. Spiga, None.

References


