

Original Research Article

DOI - 10.26479/2016.0203.11

BIOINFORMATICS ANALYSIS OF HUMAN CFTR IN COMPARISON WITH SOLVED STRUCTURES OF ABC TRANSPORTERS

Al-Zahrani, Ateeq A.

Biology and Chemistry Department, University College at Al-Qunfudah, Umm Al-Qura University, Al-Qunfudah, Saudi Arabia

ABSTRACT: Cystic fibrosis (CF) is an inherited disorder that affects about one out of every 3,000 newborns and causes a loss of transport channel activity in the cystic fibrosis transmembrane conductance regulator (CFTR) protein. Up to date, no solved structure of CFTR and this due to the difficulty of expressing, purifying and crystallizing of transmembrane proteins. In this study, structural and functional properties of solved structure ABC transporters in comparison to human CFTR were investigated using several Bioinformatics tools. Phylogenetic analysis showed that CFTR clusters with McjD, PglK and Atm1 in one lineage and they might share the same genetic origin. Alignment results revealed that NBD2 of CFTR contained several conserved amino acids (small fragments with no more 7 following residues) while NBD1 was the only completely shared domain. As a one-to-one protein comparison, ABCB1 and Pgp (known as permeability glycoproteins) showed the closest biophysical and biochemical properties to CFTR. These properties include secondary structures, sequence length, isoelectric point, GRAVY index, coiled coils, longest disorder region and the number of transmembrane helices. Based on model fitting analysis, outward facing Sav1866 and MsbA were fitted with the highest correlation coefficients (0.49 and 0.47, respectfully). A homology model of CFTR was built based on McjD, ABCB1 and Sav1866. The model showed a significant similarity to ABCB1.

KEYWORDS: Cystic fibrosis; CFTR; electron microscopy; ABC transporters; Bioinformatics, homology models of CFTR

***Corresponding Author: Prof. Dr. Al-Zahrani Ph. D**

Biology and Chemistry Department, University College at Al-Qunfudah, Umm Al-Qura University, Al-Qunfudah, Saudi Arabia

1.INTRODUCTION

ABC transporters are a superfamily of proteins that can be found in all organisms. In humans, there are 48 ABC transporters [13] and many of them involved in 13 genetic diseases such as cystic fibrosis, Sitosterolemia, adrenoleukodystrophy, and Pseudoxanthoma elasticum [40]. Structurally, ABC transporters consist of four main domains, NBD1 and 2 (nucleotide-binding domain) and TMD1 and 2 (transmembrane domain). In addition to these four domains, CFTR has an extra regulatory domain (R-domain) [35]. Up to date, the only human ABC transporter with known crystal structure is ABCB10. Most of ABC transporters in pdb (protein data bank) came from bacterial origin. CFTR is an ion channel that is encoded by the CFTR gene. Mutations in this transmembrane protein cause CF disease [9]. Many organs in the human body are affected by CF, but the lungs are considered as a more serious medical condition threatening life [31]. After 25 years of discovering the CFTR gene by Rommens [32], still there is no cure for CF. However, some trial drugs have confirmed they are effective at treating some mutations and the basic defect in the disease [39]; [43], [44]. A 3D model of the unknown protein structure (query) can be built based on known-structure protein (template). Common structural and functional features and the correct alignment between the query and template is the basis for creating a reliable 3D model. Homology modelling of protein structure is a field of structural proteomics that provide details about protein in atomic level with an accuracy comparable to that of low-resolution structures [34]. It is known that 3D structure of proteins from the same family is more conserved compared to their primary structure (sequences) [24]. Thus, a sequence similarity can be a good indication that two proteins with analogous sequence identity are much likely to have homologous structure. Several three-dimensional models of full-length CFTR (or separated domains) have been built using ABC transporters as templates [4]; [19]; [7]; [28]; [12]; [30]; [11]; [2]. All these models and others built based on transports that function as pumps which move ions against their concentration gradients. CFTR functions as an ion channel which moves ions down their concentration gradients. This difference in function between pumps and ion channels is one of some limitations regarding homology modelling of CFTR. In addition, low sequence identity of CFTR compared to ABC transporters templates (<30%) is another limitation. Beside this, the absence of R-domain in all the homology templates affects the reliability of the final 3D model structure. Despite all these limitations, homology modelling is still able to give valuable information about the structure and function of proteins [11]. The first aim of this study was to collect structural and functional information about the 10 solved structure of ABC transporters (Table 1). Comparing these data to CFTR in order to find the similarities and differences. This might help to figure out the optimal conditions for enhancing CFTR

Al-Zahrani & Ateeq RJLBPCS 2016 www.rjlbpcs.com Life Science Informatics Publications
 crystallization. The second aim was to find the best model that can be fitted into a 3D density map of CFTR generated by electron crystallography. Finally, building a 3D model of CFTR based on the previous mentioned Bioinformatics analysis.

Protein	Organism	Genbank no.	Pdb no.	AA
CFTR	<i>Homo sapiens</i>	P13569	-----	1480
SAV1866	<i>Staphylococcus aureus</i>	Q99T13	2ONJ	578
MsbA	<i>Salmonella typhimurium</i>	P63359	3B60	582
McjD	<i>Escherichia coli</i>	Q9X2W0	4PL0	580
ABCB10	<i>Homo sapiens</i>	Q9NRK6	4AYX	738
Atm1	<i>Saccharomyces cerevisiae</i>	P40416	4MYC	690
TM287–TM288	<i>Thermotoga maritima</i>	3QF4_A	4Q4A	587
PCAT1	<i>Ruminiclostridium thermocellum</i>	4RY2_A	4RY2	730
PglK	<i>Campylobacter jejuni</i>	5C78_A	5C78	564
Pgp	<i>Caenorhabditis elegans</i>	P34712	4F4C	1321
ABCB1 (Pgp in mouse)	<i>Mus musculus</i>	NP_035206	3G5U	1276

Table 1: ABC transporters involved in this study

2. MATERIALS AND METHODS

Templates Selection

The SWISS-MODEL server [5], used to search for templates matching CFTR sequence (PubMed accession code: P13569). A total of 8362 templates was found to match the target sequence. This list was filtered by a heuristic down to 50. A one template per protein was selected manually to avoid repetition. The total number of templates was 10 (Table 1).

Phylogenetic Analysis

MEGA 7 software version 7.0.14 [22] used for Phylogenetic Analysis. 11 amino acid sequences (Table 1) were aligned by Clustal W, version 1.7 [41] using the recommended amino acids settings (multiple alignment gap opening penalty 3 and multiple alignment gap extension penalty 1.8). The aligned sequences were saved as the mas format for next step analysis. Before generating a phylogenetic tree using the maximum-likelihood method, the best substitution model was computed. The LG+G+1 was the best. Several parameters were changed in the preferences dialog in Phylogeny menu as follows: For Model/Method, LG model was selected. For Phylogeny test, a bootstrap method with 200 replications was selected. For Rates among Sites, the gamma distributed with Invariants Sites (G+1) was selected. For gap/Missing Data treatment, the Partial Deletion was selected. The phylogenetic tree was constructed based on these parameters.

Alignment and conserved residues

PRALINE multiple sequence alignment, an online tool [36] used for determining conserved regions. Aligned sequences (Clustal, FASTA format) submitted to Gblocks 0.91b server [10] to predict the shared conserved domains among the 11 transporters. NCBI's conserved domain database [26] used to search for conserved domains.

Secondary Structure Predication

The secondary structure content of the solved-structure proteins obtained from the PDB (Protein Data Bank, <http://www.rcsb.org/pdb/home/home.do>). An online tool, SPIDER2 [18], used to predict CFTR secondary structure.

Assessment for crystallization

XtalPred is an online tool for prediction of protein crystallizability used to predict biochemical and biophysical features that may affect protein crystallization [38]. Another online tool, PDPredictor [3] used to create heat maps of isoelectric point (PI) and grand average hydropathy (GRAVY) of CFTR. FoldIndex, an online tool used to predict disordered regions of CFTR [29].

3D map of CFTR

The three-dimensional density map of CFTR (PDB code: 4A82, 9.0 Å resolution) generated by electron crystallography [33] used as a template for model fitting. Map-Model fitting performed by Chimera software. All the 10 transporters models (table1) were fitted into CFTR map in order to find out the best homology for CFTR. Two molecules (chain A and B) were fitted into the map by hand and then position optimized using the Fit-in-Map tool. Simulated map, resolution 9 (restricted to the same resolution as the experimental map) used to obtain the correlation index.

Homology Modeling

Modeller 9.17 software [47] used to build a 3D model of CFTR based on ABC transporter templates (ABCB1, Sav1866 and McjD). 5 scripts were modified in order to run homology modelling commands (an example of input and output files is provided with Supplementary Material). 5 models of CFTR were built. The model with the lowest value of the DOPE (Discrete Optimized Protein Energy) was selected. Further evaluation of model quality performed using RAMPAGE server (Ramachandran Plot Analysis) [25]. The created model and ABCB1 were analyzed by Pairwise Alignment tool, which is available at FATCAT server (flexible structure alignment by chaining aligned fragment pairs allowing twists) [48].

3. RESULTS AND DISCUSSION

Phylogenetic Analysis

A phylogenetic tree (Figure 1) was generated based on the maximum-likelihood method for the 11 ABC transporters (as shown in Table 1). The result shows that two main branches were created. CFTR clusters with McjD, PglK and Atm1 in one lineage. The other lineage consists of TM287–TM288, SAV1866, MsbA, ABCB10, Pgp and ABCB1. Constructing the phylogenetic tree into these two branches gives a possibility that the two groups of proteins are closely related and paralogous. Pcat1 clustered into a separate group, indicating no close sequence relationship.

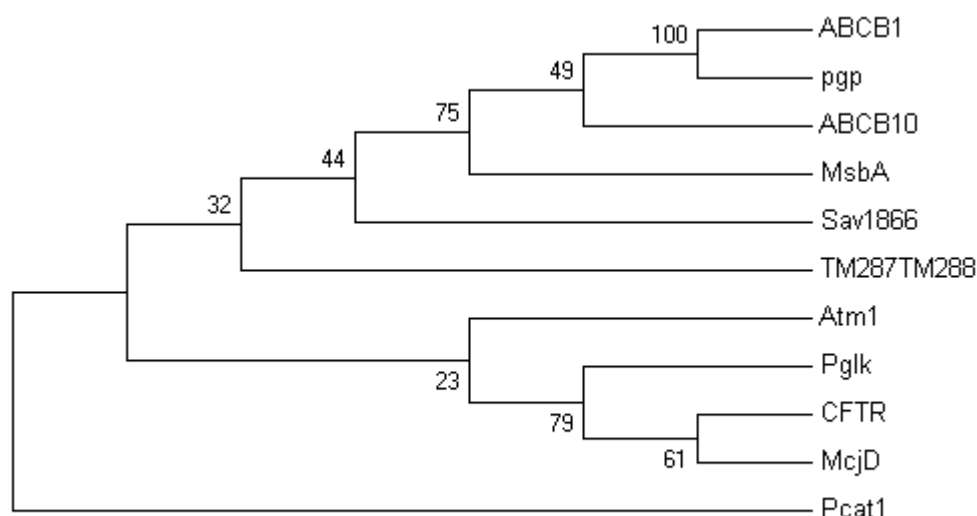


Figure 1: Molecular Phylogenetic analysis by Maximum Likelihood Method.

The evolutionary history was inferred by using the Maximum Likelihood method based on the Le_Gascuel [23]. The bootstrap consensus tree inferred from 200 replicates is taken to represent the evolutionary history of the taxa analyzed [15]. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (200 replicates) are shown next to the branches [15]. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 1.7384)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 2.4990% sites). The analysis involved 11 amino acid sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any

Alignment and Conserved regions

Alignment was performed by PRALINE multiple sequence alignment server. The results showed several conserved regions (Figure 2). Most conserved amino acids were in NBD2 of CFTR. Pairwise sequence alignment comparison (Table S1, Supplementary Material) revealed several conserved regions. The longest conserved region contained 7 residues as follows: CFTR-McJd (GSGKSTL), CFTR-ABCB10 (KILLLDE) and CFTR-Pgp (KILLLDE). ABCB10 and Pgp showed the highest Sequence identities. Gblocks 0.91b server used to predict the shared conserved domains among the 11 transporters. Figure S1 shows shared conserved domain as separate portions of ATP-binding cassette domain 1 of the cystic fibrosis transmembrane regulator. NCBI's conserved domain database used to search for conserved domains. Results are presented in Table S2.

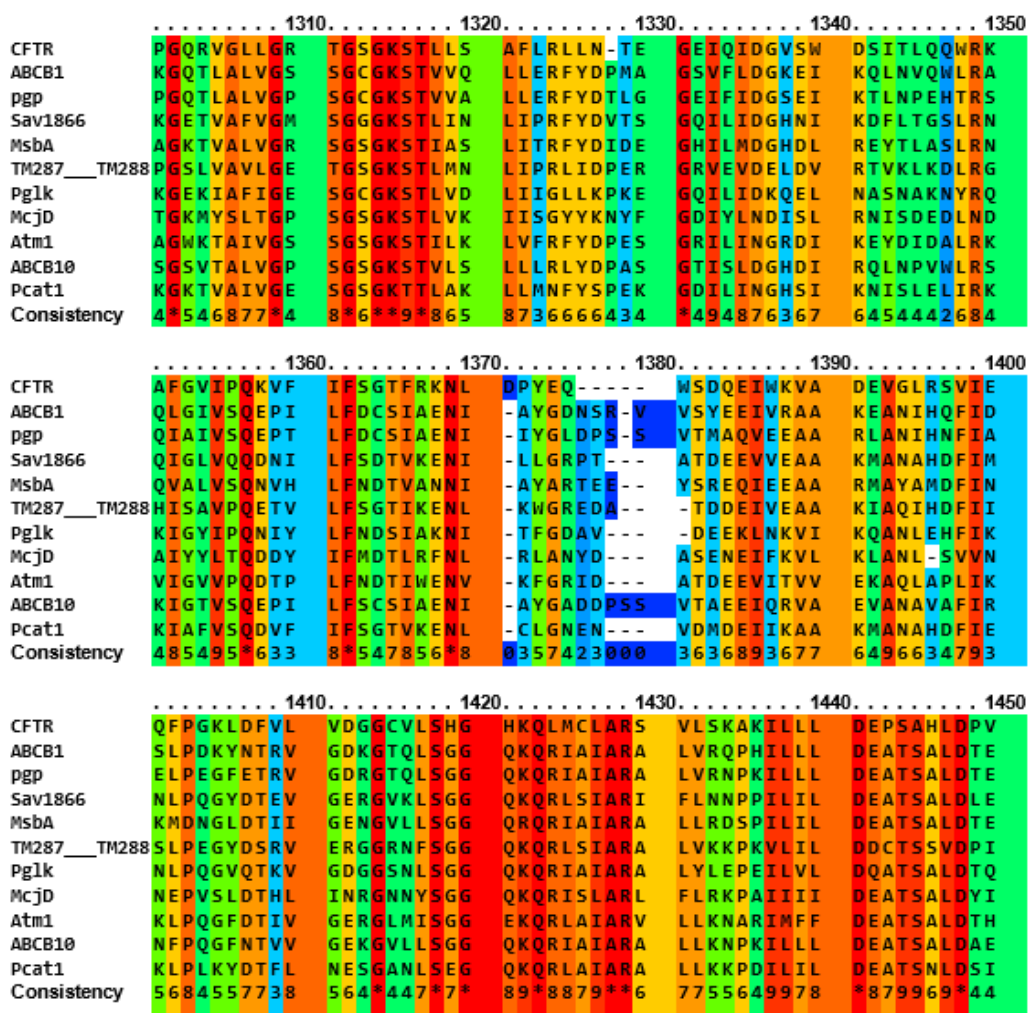


Figure 2: Alignment of the 11 transmembrane proteins. Consistency is from 1 (low conserved) to 9 (high conserved). * indicated identical amino acids.

Based on the secondary structure elements for the 11 transporters, three groups were created as shown in Table S3. CFTR, Pgp and ABCB1 formed a group of ~54% helical and ~9% beta sheet. MsbA, McjD, ABCB10 and TM287–TM288 consist of ~63% helical and ~10% beta sheet. CFTR, Pgp and ABCB1 have a higher number of helices and strands compared to the other two groups. This comparison also shows an inverse relationship between the percentage of secondary structure and the number of helices and strands. Lower percentages of the secondary structures gave a higher number of alpha helices and beta sheets. This indicates that the blue group (CFTR, Pgp and ABCB1) has more irregular secondary structural elements (random coil).

Instability index

ProtParam, online tool [17] used to compute the stability of proteins. The entire 11 transporters were predicted to be more stable with values less than 40 except of CFTR and MsbA with Instability index of 43.80 and 46.28, respectively. However, MsbA fused to a 23-residue fusion leader containing an N-terminal 10 histidine tag to enhance expression and purification [46].

Crystallizability of CFTR

XtalPred server used to predict several biochemical and biophysical features that can affect crystallizability of protein. Table S4 summarizes calculation and prediction of the 11 transporters features. ABCB1 and Pgp showed the closest features to CFTR including sequence length, iso-electric point, GRAVY index, coiled coils, longest disorder region and the number of transmembrane helices. XtalPred also provides prediction of relative crystallization probability, by two methods, crystallization class by Expert Pool (EP) and Random Forest Classifier (RF). All the 11 transporters were predicted as very difficult to be crystallized with few exceptions as shown in Table S4. PDPredictor tool used to create heat maps of pI and GRAVY of CFTR in order to identify crystalizable truncations. Figure 3 shows a wide range of pI values between 3.35 and 12.7. Most of the protein was natural (Figure 3, 1) with acidic regions, particularly in R-domain and C-terminus (dark blue color). The variation in GRAVY was between -2.8 and 3.1 with no obvious hydrophobic regions. The MCSG Z-score (the pI-GRAVY combination) reveals a possible crystallization in some regions (green color), and a difficult crystallization in other areas (dark blue) including parts of N-terminus, MSD1, NBD1, R-domain, MSD2, NBD2 and C-terminus. Further investigation was performed using a FoldIndex server in order to predict disordered regions of CFTR. The result (Figure S2) showed 11 disordered Regions, the longest was 50 residues, 268 residues in total. This result shows a good match with the MCSG Z-score in Graph 3, Figure S2. Dark blue regions, which indicate the locations of poor crystallizability.

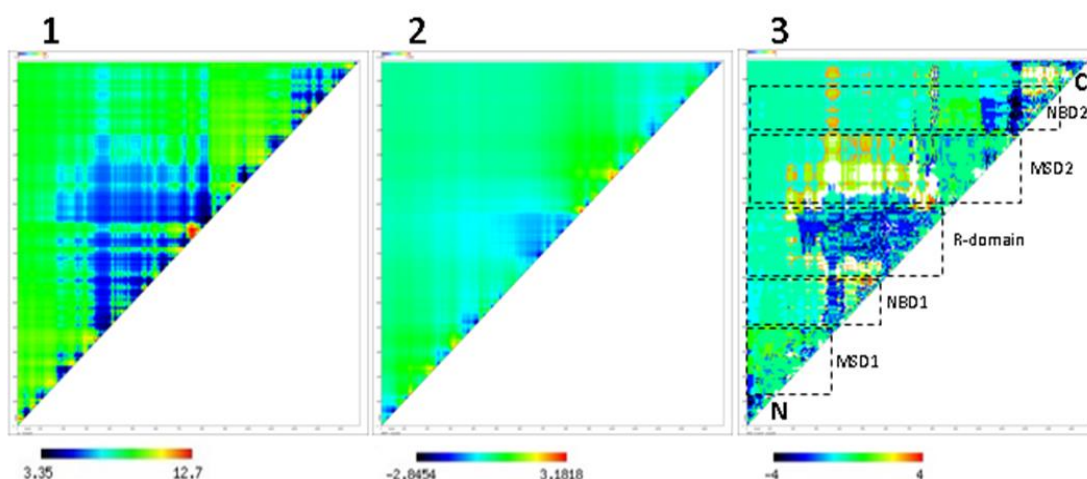


Figure 3: Three heat maps of crystallizability of full-length human CFTR. The pi (1), the GRAVY (2), and the MCSG Z-score (3). Graph 3 shows CFTR domains (dashed boxes). N is the start, C is the end of the protein.

Three-dimensional map of CFTR

Atomic models of the 10 ABC transporters were fitted into the three-dimensional density map of CFTR. Outward facing Sav1866 and MsbA gave the highest correlation coefficients with 0.49 and 0.47, respectively (Figure 4, A). In contrast, ABCB1 and Pgp with an inward facing were not able to be fitted into the map. This was due to the wide-open state of NBDs, which does not match the closed state of NBDs in the CFTR density map (Figure 4, B). However, Pgp model fitted very well into a 3D density map of CFTR with a greater separation of domain generated by single particle analysis [2].

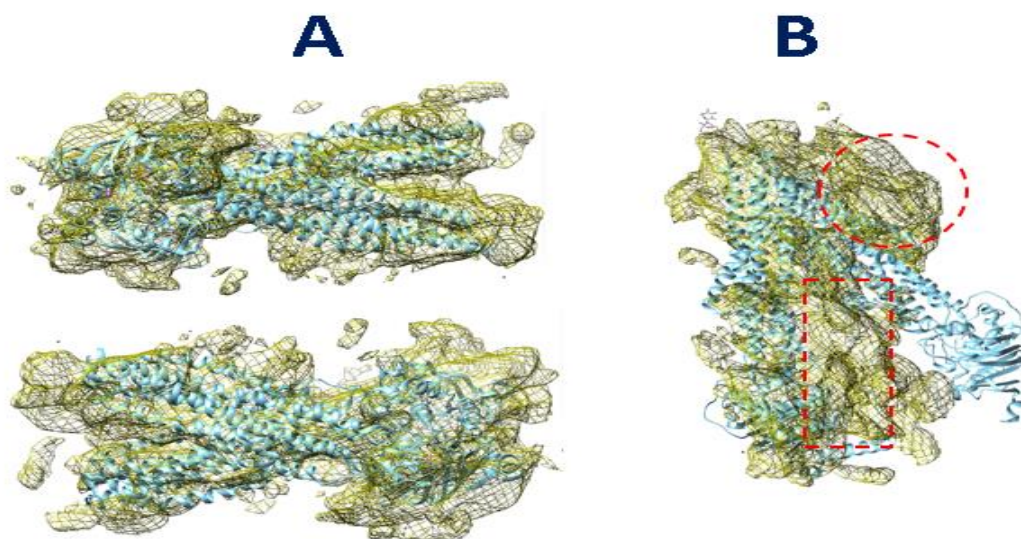


Figure 4: Fitted models in the three dimensional density map of CFTR. A, Save1866 model (blue, top) and MsbA (blue, bottom) fitted into the CFTR map (yellow). B shows the expected location of NBds (red dashed box) and the expected location of R-domain (red dashed circle).

Overview of CFTR model

A homology model of CFTR was built using Modeller 9.17 based on three ABC transporters templates that showed the best closer similarity to CFTR as revealed by the previous analysis (McjD by phylogeny analysis, ABCB1 by biophysical and biochemical properties analysis and Sav1866 by model fitting analysis). The CFTR model (Figure 5) revealed a significant similarity to ABCB1 (the best aligned template selected by Modeller run). The quality of the model evaluated by RAMPAGE server, which identified 83.2 % of residues (1230 aa) in the favored regions, 11.6 % of residues (171 aa) in the allowed regions and 5.2 % of residues (77 aa) in the outlier regions of the Ramachandran plot. The analysis of the Ramachandran plot suggested that the model is of very good quality. The FATCAT results for comparing CFTR model (query) and ABCB1 (best aligned template) showed that the two structures are significantly similar with P value < .001. The two structures alignment has 1163 equivalent positions with an RMSD (Root-mean-square deviation of atomic positions) of 2.51, with 3 twists. The best model was created using this order: ABCB1 (tseq1, best aligned template), Mcjd (tseq2, template2) and Sav1866 (tseq3, template3). New models run using Mcjd or Sav1866 as the best aligned template yielded no reasonable model. This might be a result of their short sequence length and the presence of lots of gaps in the alignment step. Solved structures NBD1 and 2 of CFTR were compared to the NBDs of CFTR model (Figure 6). The FATCAT results for comparing human NBD1 (pdb: 1XMJ) and NBD1 of CFTR model indicated that the two structures are significantly similar with P value of 2.93e-09. The two structures alignment has 182 equivalent positions with an RMSD of 2.65, without twists. Similar results were obtained for NBD2 (pdb: 3GD7) and NBD2 of CFTR model with P value of 5.55e-16 and 192 equivalent positions with an RMSD of 2.62, without twists.

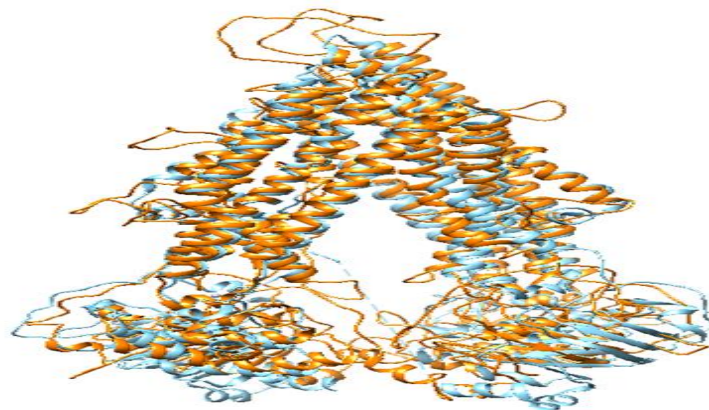








Figure 5: Superimposed model of CFTR (orange color) over ABCB1 template (light blue color). The CFTR model shows some structural position shifts in TMDs and NBDs with several disordered regions.

Legend of secondary structure icons:

 H Alpha-Helix	 T Turn
 E Extended Configuration (Beta-sheet)	 C or " " Coil
 B Isolated Beta Bridge	 G 3-10 Helix

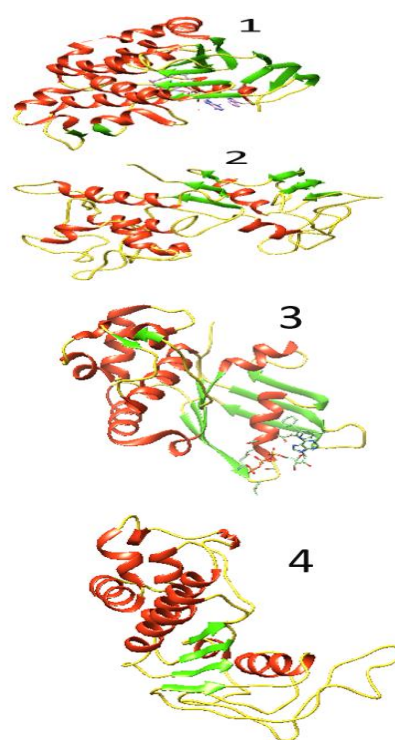
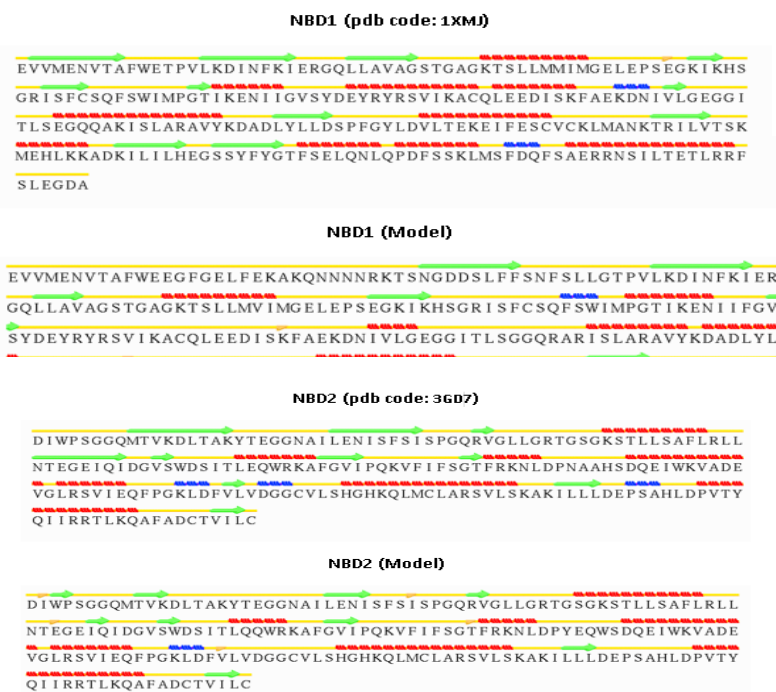


Figure 6: Secondary Structure content in NBDS of CFTR Model and Solved Structures of human NBDS. 1, human NBD1 (pdb: 1XMJ). 2, NBD1 of CFTR model. 3, human NBD2 (pdb: 3GD7). 4, NBD2 of CFTR model. The pdf files of the four structures were modified by deleting the atomic coordinates of amino acids except (391-670 aa) for NBD1 and (1202-1400 aa) for NBD2. Chimera software used to color the Secondary Structure in the four structures (1-4). STRIDE server [16] used to determine the secondary structure per residue.

DISCUSSION

Several phylogeny studies were applied to the 48 human ABC transports and divided them into seven distinct subfamilies A-G [22]; [14]; [45]. In this work, McjD (ABC-B subfamily) clustered with CFTR (ABC-C subfamily) in one group. However, a shared genetic origin of two genes, does not necessarily lead to common physiological functions. In 2002, Martinoia et al [27] found that ABC transporters in plants showed related structures with no related functions. On the other hand, apart from function, [11] suggested that the McjD atomic model could represent an open state of CFTR, even better than Sav1866 model. The analysis of biophysical and biochemical properties of the 11 ABC transporters suggested that P-glycoproteins (ABCB1 and Pgp) are the best homology for CFTR. ABCB1 and Pgp are members of the MDR/TAP subfamily (also known as multidrug resistance protein), which have a bifunctional role as pump and Cl⁻ channel [42]. CFTR functions as Cl⁻ channel and shares this activity with P-glycoproteins. Despite the similar functionality between CFTR and MDR, there was no chance to test

© 2016 Life Science Informatics Publication All rights reserved

Peer review under responsibility of Life Science Informatics Publications

2016 Sept- Oct RJLBPCS 2(3) Page No.121

PD predictor tool and FoldIndex server results revealed that the N- and C-termini of CFTR were disordered with lower possibility of crystallization. This flexibility of the two ends of CFTR was experimentally confirmed by NMR (Nuclear magnetic resonance spectroscopy) [20]; [1]; [6]. PI and GRAVY are considered as the most important biophysical properties affecting the success of protein crystallization [8]. For this reason, PD predictor tool used to create heat maps of PI and GRAVY of CFTR in order to identify crystallizable truncations. It was reported that protein with medium hydrophobicity and high acidity, increase the probability of successful crystallization [37]. Based on this study and others, it is possible to design truncation constructs of CFTR for expression and crystallization experiments avoiding the difficulty of full-length crystallization. Results in Figure 3 can be a good starting point for designing these truncations. Model fitting showed that Sav1866 yielded better fits comparing to other ABC transporters in this study. Several other studies [33]; [30]; [11] used Sav1866 as a fitting model for CFTR and that are compatible with the results of this work. Most of homology models of CFTR, which were published previously, have been based on SAV1866 and MsbA [12]. ABCB1 (and Pgp) in this study showed closest biophysical and biochemical similarity to CFTR over the other transporters. Furthermore, CFTR, ABCB1 and Pgp share approximately the same sequence length, making the produced model more reliable compared to the other templates of transporters.

4. CONCLUSION

Three parameters were examined in order to build 3D homology model of CFTR. These parameters were phylogenetic analysis, biophysical and biochemical properties comparison and fitting molecular models in an EM (electron microscopy) map. Three best CFTR homologues were predicted by the previous mentioned parameters and used to build a 3D model of CFTR. Biophysical and biochemical properties comparison suggested ABCB1 as the closest homology for CFTR. This was approved by molecular modelling which produced a 3D homology model with a significant similarity to CFTR. Furthermore, the study suggested possible crystallizable truncation constructs that can be added to the solved structure domains of CFTR (NBD1, NBD2 and R-domain).

ACKNOWLEDGEMENTS

The author wishes to thank Professor Robert C. Ford, Faculty of Life Sciences, The University of Manchester, for providing an extracted three-dimensional map of CFTR (PDB code: 4A82) and for his useful advice.

CONFLICT OF INTERESTS

The author declares no conflict of interests.

REFERENCES

- [1]Alzahrani, A. A. (2012). Structural biology of Cystic Fibrosis Transmembrane Conductance Regulator, an ATP-binding cassette protein of medical importance. Thesis (Ph.D.), University of Manchester.
- [2]Al-Zahrani, A. A., Cant, N., Kargas, V., Rimington, T., Aleksandrov, L., Riordan, J. R., Ford, R. C. (2015). "Structure of the cystic fibrosis transmembrane conductance regulator in the inward-facing conformation revealed by single particle electron microscopy". *AIMS Biophysics*. 2(2):131-152.
- [3]Babnigg, G. and A. Joachimiak (2010). "Predicting protein crystallization propensity from protein sequence." *J Struct Funct Genomics* **11**(1): 71-80.
- [4]Bianchet, M. A., Y. H. Ko, L. M. Amzel and P. L. Pedersen (1997). "Modeling of nucleotide binding domains of ABC transporter proteins based on a F1-ATPase/recA topology: structural model of the nucleotide binding domains of the cystic fibrosis transmembrane conductance regulator (CFTR)." *J Bioenerg Biomembr* **29**(5): 503-524.
- [5]Biasini, M., S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. Gallo Cassarino, M. Bertoni, L. Bordoli and T. Schwede (2014). "SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information." *Nucleic Acids Res* **42**(Web Server issue): W252-258.
- [6]Bozoky, Z., M. Krzeminski, R. Muhandiram, J. R. Birtley, A. Al-Zahrani, P. J. Thomas, R. A. Frizzell, R. C. Ford and J. D. Forman-Kay (2013). "Regulatory R region of the CFTR chloride channel is a dynamic integrator of phospho-dependent intra- and intermolecular interactions." *Proc Natl Acad Sci U S A* **110**(47): E4427-4436.
- [7]Callebaut, I., R. Eudes, J. P. Mornon and P. Lehn (2004). "Nucleotide-binding domains of human cystic fibrosis transmembrane conductance regulator: detailed sequence analysis and three-dimensional modeling of the heterodimer." *Cell Mol Life Sci* **61**(2): 230-242.
- [8]Canaves, J. M., R. Page, I. A. Wilson and R. C. Stevens (2004). "Protein biophysical properties that correlate with crystallization success in *Thermotoga maritima*: maximum clustering strategy for structural genomics." *J Mol Biol* **344**(4): 977-991.
- [9]Cant, N., N. Pollock and R. C. Ford (2014). "CFTR structure and cystic fibrosis." *Int J Biochem Cell Biol* **52**: 15-25.
- [10]Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." *Mol Biol Evol* **17**(4): 540-552.
- [11]Corradi, V., P. Vergani and D. P. Tieleman (2015). "Cystic Fibrosis Transmembrane Conductance Regulator (CFTR): CLOSED AND OPEN STATE CHANNEL MODELS." *J Biol Chem* **290**(38): 22891-22906.

- Al-Zahrani & Ateeq RJLBPCS 2016 www.rjlbpcs.com Life Science Informatics Publications
- [12]Dalton, J., O. Kalid, M. Schushan, N. Ben-Tal and J. Villà-Freixa (2012). "New model of cystic fibrosis transmembrane conductance regulator proposes active channel-like conformation." *J Chem Inf Model* **52**(7): 1842-1853.
- [13]Dean, M., Y. Hamon and G. Chimini (2001). "The human ATP-binding cassette (ABC) transporter superfamily." *J Lipid Res* **42**(7): 1007-1017.
- [14]Dean, M., A. Rzhetsky and R. Allikmets (2001). "The human ATP-binding cassette (ABC) transporter superfamily." *Genome Res* **11**(7): 1156-1166.
- [15]Felsenstein J. (1985). "Confidence limits on phylogenies: An approach using the bootstrap". *Evolution*. Vol. 39, No. 4. pp. 783-791.
- [16]Frishman, D. and P. Argos (1995). "Knowledge-based protein secondary structure assignment." *Proteins* **23**(4): 566-579.
- [17]Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A;. (2005) "Protein Identification and Analysis Tools on the ExPASy Server". (In) John M. Walker (ed): *The Proteomics Protocols Handbook*, Humana Press. pp. 571-607.
- [18]Heffernan, R., K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang and Y. Zhou (2015). "Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning." *Sci Rep* **5**: 11476.
- [19]Hoedemaeker, F. J., A. R. Davidson and D. R. Rose (1998). "A model for the nucleotide-binding domains of ABC transporters based on the large domain of aspartate aminotransferase." *Proteins* **30**(3): 275-286.
- [20]Ivancic, M., A. M. Spuches, E. C. Guth, M. A. Daugherty, D. E. Wilcox and B. A. Lyons (2005). "Backbone nuclear relaxation characteristics and calorimetric investigation of the human Grb7-SH2/erbB2 peptide complex." *Protein Sci* **14**(6): 1556-1569.
- [21]Klein, I., B. Sarkadi and A. Váradi (1999). "An inventory of the human ABC proteins." *Biochim Biophys Acta* **1461**(2): 237-262.
- [22]Kumar, S., G. Stecher and K. Tamura (2016). "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets." *Mol Biol Evol* **33**(7): 1870-1874.
- [23]Le, S. Q. and O. Gascuel (2008). "An improved general amino acid replacement matrix." *Mol Biol Evol* **25**(7): 1307-1320.
- [24]Lesk, A. M. and Chothia, C. (1980) "How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins". *J. Mol. Biol.* 136, 225–270.
- [25]Lovell, S. C., I. W. Davis, W. B. Arendall, P. I. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson and D. C. Richardson (2003). "Structure validation by C α geometry: phi,psi and C β deviation." *Proteins* **50**(3): 437-450.

[26]Marchler-Bauer, A., M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng and S. H. Bryant (2015). "CDD: NCBI's conserved domain database." *Nucleic Acids Res* **43**(Database issue): D222-226.

[27]Martinoia, E., M. Klein, M. Geisler, L. Bovet, C. Forestier, U. Kolukisaoglu, B. Müller-Röber and B. Schulz (2002). "Multifunctionality of plant ABC transporters--more than just detoxifiers." *Planta* **214**(3): 345-355.

[28]Mornon, J. P., P. Lehn and I. Callebaut (2008). "Atomic model of human cystic fibrosis transmembrane conductance regulator: membrane-spanning domains and coupling interfaces." *Cell Mol Life Sci* **65**(16): 2594-2612.

[29]Prilusky, J., C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman and J. L. Sussman (2005). "FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded." *Bioinformatics* **21**(16): 3435-3438.

[30]Rahman, K. S., G. Cui, S. C. Harvey and N. A. McCarty (2013). "Modeling the conformational changes underlying channel opening in CFTR." *PLoS One* **8**(9): e74574.

[31]Riordan, J. R. (2008). "CFTR function and prospects for therapy." *Annu Rev Biochem* **77**: 701-726.

[32]Rommens, J. M., M. C. Iannuzzi, B. Kerem, M. L. Drumm, G. Melmer, M. Dean, R. Rozmahel, J. L. Cole, D. Kennedy and N. Hidaka (1989). "Identification of the cystic fibrosis gene: chromosome walking and jumping." *Science* **245**(4922): 1059-1065.

[33]Rosenberg, M. F., L. P. O'Ryan, G. Hughes, Z. Zhao, L. A. Aleksandrov, J. R. Riordan and R. C. Ford (2011). "The cystic fibrosis transmembrane conductance regulator (CFTR): three-dimensional structure and localization of a channel gate." *J Biol Chem* **286**(49): 42647-42654.

[34]Sánchez, R. and Sali, A. (1997) "Advances in comparative protein-structure modeling". *Curr. Opin. Struct. Biol.* **7**, 206–214.

[35]Seeger, M, Enrica Bordignon and Michael Hohl. (2016) "ABC Exporters from a Structural Perspective". *ABC Transporters - 40 Years on*. Springer. pp.65-84.

[36]Simossis, V. A. and J. Heringa (2005). "PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information." *Nucleic Acids Res* **33**(Web Server issue): W289-294.

[37]Slabinski, L., L. Jaroszewski, A. P. Rodrigues, L. Rychlewski, I. A. Wilson, S. A. Lesley and A. Godzik (2007). "The challenge of protein structure determination--lessons from structural genomics." *Protein Sci* **16**(11): 2472-2482.

[38]Slabinski, L., L. Jaroszewski, L. Rychlewski, I. A. Wilson, S. A. Lesley and A. Godzik (2007). "XtalPred: a web server for prediction of protein crystallizability." *Bioinformatics* **23**(24): 3403-3405.

[39]Sosnay, P. R., K. R. Siklosi, F. Van Goor, K. Kaniecki, H. Yu, N. Sharma, A. S. Ramalho, M. D. Amaral, R. Dorfman, J. Zielenski, D. L. Masica, R. Karchin, L. Millen, P. J. Thomas, G. P. Patrinos,

Al-Zahrani & Ateeq RJBPCS 2016 www.rjlbpcs.com Life Science Informatics Publications
M. Corey, M. H. Lewis, J. M. Rommens, C. Castellani, C. M. Penland and G. R. Cutting (2013).
"Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator
gene." *Nat Genet* **45**(10): 1160-1167.

[40]Stefková, J., R. Poledne and J. A. Hubáček (2004). "ATP-binding cassette (ABC) transporters in
human metabolism and diseases." *Physiol Res* **53**(3): 235-243.

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

[42]Valverde, M. A., M. Díaz, F. V. Sepúlveda, D. R. Gill, S. C. Hyde and C. F. Higgins (1992).
"Volume-regulated chloride channels associated with the human multidrug-resistance P-
glycoprotein." *Nature* **355**(6363): 830-833.

[43]Van Goor, F., S. Hadida, P. D. Grootenhuis, B. Burton, D. Cao, T. Neuberger, A. Turnbull, A.
Singh, J. Joubran, A. Hazlewood, J. Zhou, J. McCartney, V. Arumugam, C. Decker, J. Yang, C. Young,
E. R. Olson, J. J. Wine, R. A. Frizzell, M. Ashlock and P. Negulescu (2009). "Rescue of CF airway
epithelial cell function in vitro by a CFTR potentiator, VX-770." *Proc Natl Acad Sci U S A* **106**(44):
18825-18830.

[44]Van Goor, F., S. Hadida, P. D. Grootenhuis, B. Burton, J. H. Stack, K. S. Straley, C. J. Decker, M.
Miller, J. McCartney, E. R. Olson, J. J. Wine, R. A. Frizzell, M. Ashlock and P. A. Negulescu (2011).
"Correction of the F508del-CFTR protein processing defect in vitro by the investigational drug VX-
809." *Proc Natl Acad Sci U S A* **108**(46): 18843-18848.

[45]Vasiliou, V., K. Vasiliou and D. W. Nebert (2009). "Human ATP-binding cassette (ABC)
transporter family." *Hum Genomics* **3**(3): 281-290.

[46]Ward, A., C. L. Reyes, J. Yu, C. B. Roth and G. Chang (2007). "Flexibility in the ABC transporter
MsbA: Alternating access with a twist." *Proc Natl Acad Sci U S A* **104**(48): 19005-19010.

[47]Webb, B. and A. Sali (2016). "Comparative Protein Structure Modeling Using MODELLER." *Curr
Protoc Bioinformatics* **54**: 5.6.1-5.6.37.

[48]Ye, Y. and A. Godzik (2003). "Flexible structure alignment by chaining aligned fragment pairs
allowing twists." *Bioinformatics* **19 Suppl 2**: ii246-255.

Supplementary Material

Aligned Proteins	1 aa	2 aa	3aa	4aa	5aa	6aa	7aa	Sequence Identities
CFTR-SAV1866	59	9	4	0	1	0	0	100
CFTR-MsbA	66	16	2	0	1	0	0	113
CFTR-McjD	55	14	2	0	0	0	1	102
CFTR-ABCB1	135	28	5	1	1	1	0	229
CFTR-ABCB10	57	14	2	1		1	1	117
CFTR-Atm1	74	18	0	0	0	1	0	118
CFTR-TM287– TM288	47	22	3	0	0	1	0	120
CFTR-PCAT1	72	19	3	2	0	0	0	129
CFTR-PglK	44	14	2	0	1	0	0	89
CFTR-Pgp	126	22	8	1	1	0	1	216

Table S1: Pairwise sequence alignment comparison shows identical amino acids.

Xaa shows the number of amino acids in the conserved region. CFTR-ABCB1 and Pgp recorded the highest sequence identities. Pairwise alignment was performed by PRALINE.

protein	Accession	Domain Description	Interval	E-Value
CFTR	cd03291	ATP-binding cassette domain 1	389-670	0e+00
	cd03289	ATP-binding cassette domain 2	1208-1480	0e+00
	pfam14396	regulator domain	639-850	1.54e-137
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	83-350	7.07e-41
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	862-1147	3.72e-49
Sav1866	cd03251	ATP-binding cassette domain	340-573	3.04e-137
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	12-294	4.35e-39
MsbA	cd03251	ATP-binding cassette domain	342-576	2.61e-149
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	27-298	2.32e-54
McdJ	cd03228	ATP-binding cassette domain	345-554	7.83e-56
	cl00549	ABC transporter transmembrane region, a unit of six transmembrane	48-300	2.64e-04
ABCB1	cd03249	ATP-binding cassette domain 1	389-625	2.71e-151
	cd03249	ATP-binding cassette domain 2	1031-1270	1.09e-151
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	50-341	3.14e-72
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	711-987	1.50e-62
ABCB10	cl02554	The PWWP domain, named for a conserved Pro-Trp-Trp-Pro motif	1-26	6.86e-03
	cd03249	ATP-binding cassette domain	493-733	5.27e-142
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	171-441	1.17e-48
Atm1	cd03253	ATP-binding cassette domain	436-674	1.42e-147
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	111-389	4.08e-39
TM287-TM288	cl21455	P-loop containing Nucleoside Triphosphate Hydrolases	340-566	3.80e-90
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	23-296	2.56e-41
PCAT1	cd02418	A sub-family of peptidase family C39	13-146	2.48e-58
	cl21455	P-loop containing Nucleoside Triphosphate Hydrolases	489-721	5.80e-106
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	172-442	7.56e-40
PgIK	cl21455	P-loop containing Nucleoside Triphosphate Hydrolases	353-563	4.75e-77
	cl00549	ABC transporter transmembrane region, a unit of six transmembrane	16-250	5.51e-07
Pgp	cd03249	ATP-binding cassette domain 1	416-653	3.34e-146
	cd03249	ATP-binding cassette domain 2	1077-1316	3.83e-148
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	77-369	7.50e-56
	pfam00664	ABC transporter transmembrane region, a unit of six transmembrane	754-1031	8.43e-50

Table S2: Conserved domains located into the 11 transporters.

	Protein	% helical	% beta sheet	No. of helices	No. of strands
1	CFTR	54	9	67	35
	Pgp	54	9	50	30
	ABCB1	55	7	43	24
2	MsbA	63	10	24	14
	McjD	62	10	20	14
	ABCB10	62	10	22	14
	TM287– TM288	62	10	25	14
3	SAV1866	63	9	32	12
	Atm1	59	10	29	14
	PCAT1	57	10	28	23
	PglK	61	6	19	12

Table S3: Secondary structure content of the 11 ABC transporters. (1) Indicates a group of proteins sharing a similar secondary structure content (including CFTR). (2) Indicates a group categorized according to its secondary structure similarities. (3) Boxes indicate the rest of the transports that have a different secondary structure percentage.

Target id	EP-class	RF-class	Length	Gravy index	Instability index (II)	Isoelectric point (pI)	Coiled coils	Longest disorder region	Percentage of coil structure	Transmembrane helices (TM)	Signal peptides (SP)
<u>TM287-TM288</u>	5	11	587	0.22	25.25	9.07	0	19	24	6	No
<u>ABCB1</u>	5	11	1276	0.03	31.65	8.94	21	65	27	11	No
<u>Pcat1</u>	5	11	730	0.20	24.81	8.79	21	18	27	5	No
<u>pgp</u>	5	11	1321	0.04	31.64	7.99	45	81	29	11	No
<u>McjD</u>	5	11	580	0.22	36.65	8.09	21	20	24	5	47
<u>Pglk</u>	5	5	564	0.21	30.66	9.26	0	19	22	6	45
<u>MsbA</u>	5	11	582	0.14	46.31	8.32	0	19	23	5	No
<u>ABCB10</u>	5	11	738	0.11	38.19	9.87	0	86	34	5	No
<u>Sav1866</u>	5	6	578	0.20	30.68	6.72	21	18	24	5	No
<u>Atm1</u>	5	10	690	0.005	28.84	9.62	21	50	29	5	No
<u>CFTR</u>	5	11	1480	0.02	43.83	8.91	21	73	31	11	No

Table S4: Biochemical and biophysical features that may affect crystallizability of the 11 transporters.

Yellow highlighted numbers indicate the closest values to CFTR. The analysis was performed by XtalPred server.

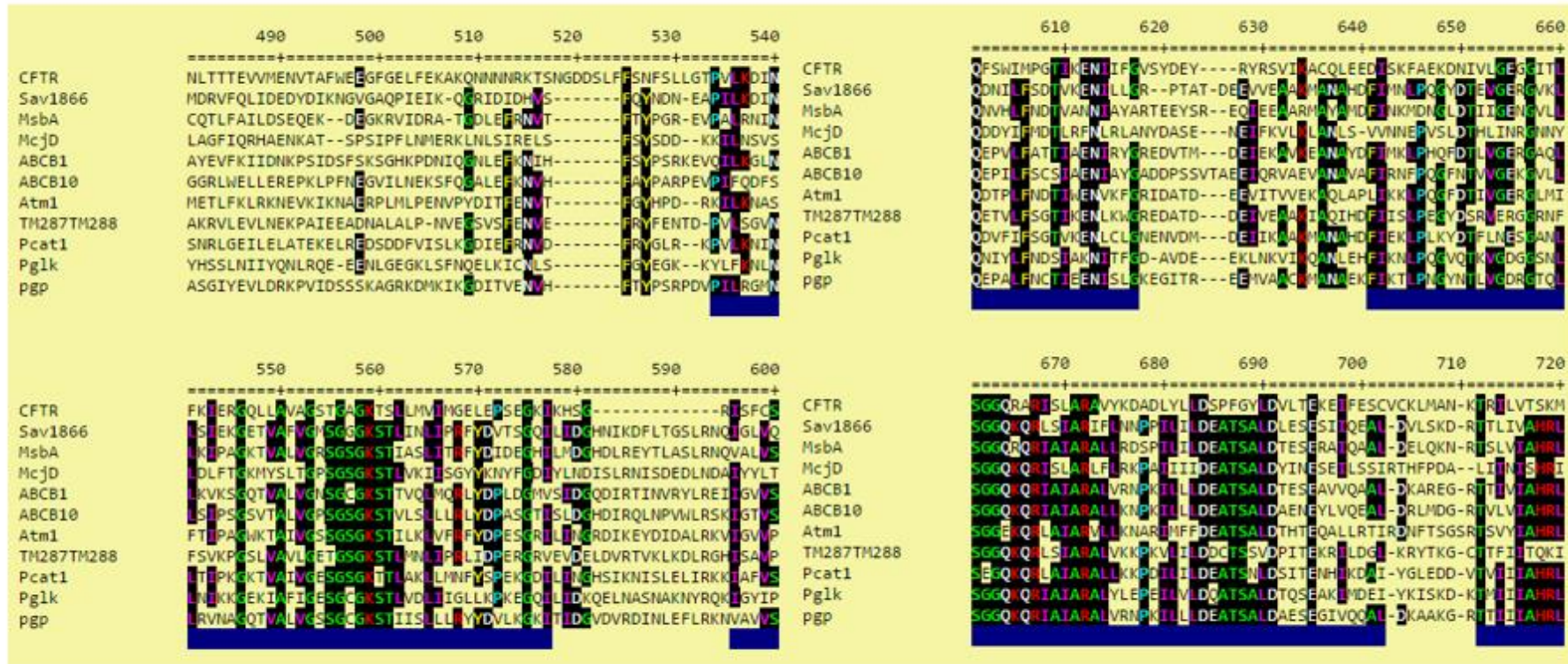


Figure S1: A portion of alignment results reveals conserved domain among the 11 transporters. Thick dark blue color indicates a conserved domain, which is mostly located in the NBD1 region of CFTR.

1	MQRSPLEKAS	VVSKLFFSWT	RPILRKGYSR	RLELSDIYQI	PSVDSADNLS
51	EKLEREWDR	LASKKNPKLI	NALRRCFWR	FMFYGIFLYL	GEVTKAVQPL
101	LLGRIIASYD	PDNKEERSIA	IYLGIGLCLL	FIVRTLHHP	AIFGLHHIGM
151	QMRIAMFSLI	YKKTLLKSSR	VLDKISIGQL	VSLLSNNLNK	FDEGLALAHF
201	WVIAPLQVAL	LMGLIWELLQ	ASAF CGLGFL	IVLALFQAGL	GRMMMKYRDQ
251	RAGKISERLV	ITSEMIENIQ	SVKAYCWEEA	MEKMIENLRQ	TELKLTRKAA
301	YVRYFNSSAF	FFSGFFVWFL	SVLPYALIKG	IILRKIFTTI	SFCIVLRMAV
351	TRQFPWAVQT	WYDSLGAINK	IQDFLQKQEY	KTLEYNLTTT	EVVMENVTAF
401	WEEGFGELE	KAKQNNNRK	TSNGDDSLFF	SNFSLLGTPV	LKDINFKIER
451	GQLLAVAGST	GAGKTSLLMV	IMGELEPSEG	KIKHSGRISF	CSQFSWIMPG
501	TIKENIIFGV	SYDEYRYSV	IKACQLEEDI	SKFAEKDNIV	LGGEGITLSG
551	GQRARISLAR	AVYKDADLYL	LDSPPGYLDV	LTEKEIFESC	VCKLMANKTR
601	ILVTSKMEHL	KKADKILILH	EGSSYFYGTF	SELQNLQPDF	SSKLMGCDSE
651	DQFSAERRNS	ILTEHLRFS	LEGDAPVSWT	ETKKQSFQKT	GEFGEKRKNS
701	ILNPINSIRK	FSIVQKTPLQ	MNGIEEDSDE	PLERRLSLVP	DSEQGEAILP
751	RISVISTGPT	LQARRRQSVL	NLMTHSVNQG	QNIHRKTTAS	TRKVS LAPQA
801	NLTELDIYSR	RLSQETGLEI	SEEINEEDLK	ECFFDDMESI	PAVTTWNTYL
851	RYITVHKSLI	FVLWCLVIF	LAEVAASLVV	LWLLGNTPLQ	DKGNSTHSRN
901	NSYAVIITST	SSYVFIYIV	GVADTLLAMG	FFRGLPLVHT	LITVSKILHH
951	KMLHSVLQAP	MSTLNTLKAG	GILNRF SKDI	AILDDLPLT	IFDFIQLLLI
1001	VIGAIAVVAV	LQPYIFVATV	PVIVAFIMLR	AYFLQTSQQL	KQLESEGRSP
1051	IFTHLVTSLK	GLWTLRAFGR	QPYFETLFHK	ALNLHTANWF	LYLSTLRWFQ
1101	MRIEMIFVIF	FIAVTFISIL	TTGEGEGRVG	IILTLAMNIM	STLQWAVNSS
1151	IDVDLSMRSV	SRVFKFIDMP	TEGKPTKSTK	PYKNGQLSKV	MIENSHVKK
1201	DDIWPSSGQM	TVKDLTAKYT	EGGNAILENI	SFSISPGQRV	GLLGRGTSGK
1251	STLLSAFLRL	LNTEGEIQID	GVSWDSITLQ	QWRKAFGVIP	QKVFIFSGTF
1301	RKNLDPYEQW	SDQEIWKVAD	EVGLRSVIEQ	FPGKLDVFLV	DGGCVLSHGH
1351	KQLMCLARSV	LKAKILLLD	EPSAHLDPVT	YQIIRRTLKQ	AFADCTVILC
1401	EHRIEAMLEC	QQFLVIEENK	VRQYDSIQKL	LNERSLFRQA	ISPSDRVKLF
1451	PHRNSSKCKS	KPQIAALKEE	TEEEVQDTRL		

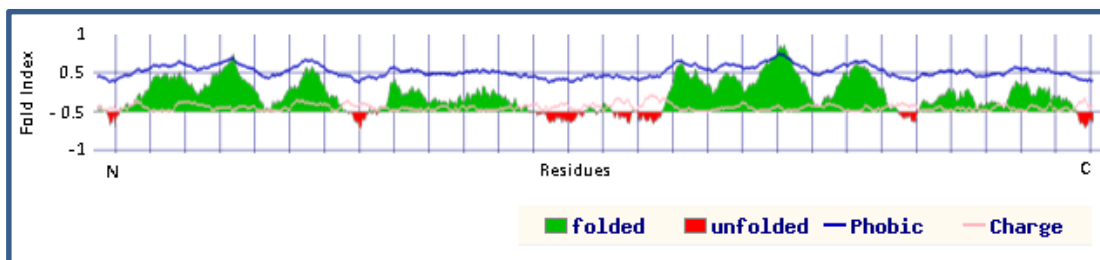


Figure S2:

Disordered regions of CFTR. Upper graph shows the locations of folded and unfolded regions of CFTR sequence.

Bottom graph shows the Fold Index for CFTR. Unfold ability 0.236 (Charge: 0.013, Phobic: 0.503).