Research Journal of Biological Sciences 5 (6): 404-413, 2010
ISSN: 1815-8846

Protein Function Extrapolation via Inventorical Clustering and Predictive Comparative Analysis of Sequence Structure Function Relationships of a *Burkholderia pseudomallei* ATP Binding Protein

Mohd Firdaus-Raih, Roohaida Othman and Rahmah Mohamed
School of Biosciences and Biotechnology, Faculty of Science and Technology,
University of Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, Malaysia

Abstract: The tropical soil pathogen *Burkholderia pseudomallei* is known to secrete a variety of extracellular virulence factors and is resistant to a wide range of antibiotics. ATP Binding Cassette (ABC) transport systems typically consist of three proteins encoded by genes in the same operon or neighboring operons. This functionally diverse protein super family typically carry out vector driven transport across cell membranes. The studies have identified an open reading frame that was predicted to code for an ATP binding protein in *B. pseudomallei*. A whole genome inventorization and function classification by sequence similarity clustering was done by comparison of ABC transporters from three completed genomes. The predicted protein sequence of the ATP binding domain was successfully modelled onto the crystal structure of an ATP binding subunit for the histidine permease of *Salmonella typhimurium*. The putative ATP binding site was identified and the model deemed to be functionally viable from a structural-mechanics point of view. We have also detected two highly conserved residues which appear to be independent of the known motifs associated with the ATP binding subunit. This integrated process was able to identify and infer function to an ABC domain from *B. pseudomallei* despite low target template sequence identity (26.4%). This approach enabled ABC transporters with varying functions to be classified predictively.

Key words: Burkholderia pseudomallei, ATP binding transporter, structure function relationships, predictively, successfully, super family

INTRODUCTION

Burkholderia pseudomallei is the causative agent of melioidosis, a potentially fatal disease that affects humans as well as livestock. This gram negative pathogen is endemic to Southeast Asia and Northern Australia, however melioidosis and *B. pseudomallei* have been reported in India, Africa, Europe, Latin America and Southern China (Dance, 2002). The pathogenicity of this bacterium may involve several extracellular enzymes, including haemolysin, siderophore, exotoxin and proteases (Brett and Woods, 2000). *B. pseudomallei* is also resistant to a wide range of antibiotics (Dance, 2002). The ATP Binding Cassette (ABC) is the largest protein family known (Henikoff *et al*., 1997).

ABC transporter proteins are an important family of proteins which is believed to play a significant role in the secretion of a probable Serratia like metalloprotease and possible antibiotic resistance in *B. pseudomallei*. The ABC transporters have been shown to be involved in virulence contributing secretion of proteases in *Pseudomonas aeruginosa* (alkaline protease), *Serratia marcescens* (metalloprotease) and *Erwinia* sp. (metalloprotease) (Hase and Finkelstein, 1993). Multi drug resistance has also been attributed to this family of proteins (Putman *et al*., 2000) and may be an important drug efflux pathway which contributes to antibiotic resistance in *B. pseudomallei*.

The ABC transporters are a super family of prokaryotic and eukaryotic proteins involved in the ATP driven vectorial movement of a single or a group of related substrates such as ions, sugars, amino acids, oligosaccharides as well as peptides and proteins that lack a classical N-terminal signal peptide (Wandersman, 1998). The typical feature of these ABC transporters is the presence of a large and conserved hydrophilic domain bearing the ATP binding site called the ATP binding cassette or nucleotide binding domain (Wandersman, 1998).

ABC transporters in both prokaryotes and eukaryotes are believed to consist of four domains, two cytoplasmic ABC domains and two hydrophobic membrane spanning

Corresponding Author: Mohd Firdaus-Raih, School of Biosciences and Biotechnology, University of Kebangsaan Malaysia, Malaysia

domains (Wandersman, 1998). ABC exporters of gram negative bacteria comprise of two helper proteins in addition to the ABC protein, the first, a membrane fusion protein is anchored to the inner membrane; the second is an outer membrane protein (Wandersman, 1998). Sequence alignments by Walker *et al*. (1982) revealed the presence of two conserved Nucleotide Binding Domains (NBD) (Walker *et al*., 1982).

The consensus sequence for the first Walker motif (Walker A) is GX(S/T)GXGK(S/T). This motif occurs in known structures as a glycine rich loop and has also been shown to interact with the γ-phosphate of ATP or GTP in the presence of a Mg^{2+} ion in several structures (Matte *et al*., 1996).

The exact mode of interaction with ATP for the ILLLD sequence of the Walker B motif is still unclear due to the absence of conclusive explanations. However, mutations of this motif in DNA heli caseII have been shown to affect ATP hydrolysis (Brosh and Matson, 1995). By taking advantage of the presence of highly conserved motifs, the structural studies of proteins from the ABC super family can be used as models for further understanding of other similar transmembrane transport systems in the same or similar organism. The ABC protein super family was a useful model for the research as the sequences involved were identifiably homologous despite the existence of divergent sequences and at times very low sequence identities.

We have predicted a tertiary structure model for a novel ATP binding domain consisting of 258 amino acids which to the knowledge are the first structural study reported for this family of proteins in *B. pseudomallei* in general and *B. pseudomallei* D286 specifically. The involvement of the ABC protein super family in a diverse range of biological functions including possible virulence contributing and drug resistance pathways, resulted in the need to inventorize and study the structure and function of the ATP binding cassette as an initial step to create a model pathway for future studies of transmembrane transport via ABC transporters in *B. pseudomallei*.

The researach further demonstrates the usefulness of presently available computational protein sequence structure function determination methods, especially when used in tandem for function determination despite low sequence identities and limited template structures. A comparative study was also done to identify, functionally classify and inventorize ABC transporters in the genome of the completely sequenced *B. pseudomallei* K96243. The methods exercised, mainly utilised either academically or publicly available software and database resources to achieve a refined and conclusive extrapolation of structure and function from a genomic DNA sequence.

MATERIALS AND METHODS

Sequence analysis: The DNA fragment carrying the nucleotide sequence (GenBank accession numbers: nucleotide = AF390557; protein = AAK67358) was fished out of a *B. pseudomallei* D286 clinical isolate genomic library and putative coding regions were predicted using hidden markov model profiles of completed bacterial genomes specifically those of *Escherichia coli* and *P. aeruginosa* via GeneMark.hmm. The raw sequences and predicted coding regions were subjected to database searching using the NCBI BLAST family of programs (Altschul *et al*., 1997) on the NCBI (Bethesda, MD, USA) WWW server. Initial database searches by translated (BLASTX) and protein (BLASTP) BLAST searches were followed by PSI-BLAST searches.

The relevant 20 closely related non-redundant protein sequences were used to create a multiple alignment profile to detect sequence conservation and for phylogenetic analysis using ClustalW (Thompson *et al*., 1994) and PHYLIP (Lim and Zhang, 1999). Hydropathy profiles were generated using the Kyte and Doolittle (1982)'s method while transmembrane region predictions were done using the web servers MEMSAT-2 (Jones *et al*., 1994) and DAS (Cserzo *et al*., 1997). Inventorization and sequence alignments based functional classification.

Sequences of completely annotated and sequenced bacterial genomes as well as sequences from completely sequenced genomes with functional annotation in progress was used for this phase of the study. The PFAM domain search was limited to ABC transporter while *Escherichia coli* and *Ralstonia solanacearum* were the two chosen completed bacterial genomes. *E. coli* was chosen on the basis that it was a well known and well characterized genome while *R. solanacearum* was chosen because of a possible close phylogenetic relationship to *B. pseudomallei* based on observations from BLAST searches against the GenBank database. Multiple sequence alignments were done for the sequences retrieved. Alignments were done for the ABC transporter sequences of *B. pseudomallei*, *E. coli* and *R. solanacearum*.

Phylogenetic trees were built from these alignments using PHYLIP. A further alignment of all these ABC transporter sequences were also done. The alignments and resulting trees were annotated and the function of specific groups, families and subfamilies of these inventorized ABC transporters were identified.

Structure prediction: The sequence was further submitted to 6 automatic fold prediction servers. Total 4 of these (3D-PSSM, SAM-T99, UCLA-DOE Fold Recognition and GenThreader) were chosen for their

success rate in the evaluation of automatic fold prediction ability. 3D-PSSM uses structural alignment to aid fold recognition for which the result is stated to be highly confident if the e-value is <0.05 (Karplus *et al*., 1999).

SAM-T99 uses Hidden Markov Models (HMM) and the score results from a comparison of the result with that obtained using a reversed sequence with more negative scores being more reliable (<1% of false positives from a score of <-28) (Fischer and Eisenberg, 1996). UCLA-DOE Fold Recognition uses a Z-score computed from the distribution of raw scores (Jones, 1999). GenThreader uses a neural network to combine various aspects of sequence structure alignment and produces a score between 0 and 1 with 1 being 100% certainty of the correct prediction and <0.5 unlikely to be correct (Fischer, 2000).

Bioinbgu uses a hybrid method combining sequence derived properties with evolutionary information (Shi *et al*., 2001). Fugue uses structural environment specific substitution tables and structure dependent gap penalties where scores for amino acid matching and insertions/deletions are evaluated depending on the local environment of each amino acid residue in a known structure (Berman *et al*., 2000).

Comparative modeling and model refinement: The structure that gave an optimal overall sequence alignment and agreed upon by all the fold recognition methods used was chosen as the template. The consensus structure, an ATP binding subunit of a histidine permease (PDB entry: 1B0U) from *Salmonella typhimurium* solved by X-ray crystallography to a resolution of 1.5Å (Hung *et al*., 1998) was downloaded from the protein databank (Sali and Blundell, 1993). Alignments were done for the predicted sequence with twenty of the best protein hits as well as the template structure sequence (1B0U) using ClustalW. The template sequence and predicted ABC transporter sequence (264 amino acids) were loaded into the insightII environment (MSI, ver. 98, CA).

The Homology module was used to create an optimal alignment between the two sequences; the predicted sequence was edited to 258 amino acids to match the template. Initial models were then generated by satisfying spatial restraints using modeller (Sali and Blundell, 1993).

Short contacts if any were removed by manually rotating the side chains. The model was refined by removing unfavourable conformations with a combination of automated rotamer searches using the Homology module and energy minimizations with the CHARMm force field. Models were examined and validated using Procheck (Laskoswki *et al*., 1993), Errat (Colovos and Yeates, 1993) and Verify3D (Luthy *et al*.,1992).

RESULTS AND DISCUSSION

Sequence analysis: Screening of a genomic library by PCR for ABC transporter genes using degenerate primers based on the ATP binding cassette proteins of *P. aeruginosa* (AprD) was followed by sequencing of the amplified fragments. A coding region for a probable ATP binding domain belonging to an ABC transporter protein family from *B. pseudomallei* was identified via Hidden Markov model profiles. The predicted gene and translated protein sequence was deposited in GenBank (GenBank accession numbers: nucleotide = AF390557; protein = AAK67358) and will be referred by the accession numbers or as BpABC1.

The DNA sequence was originally derived from a genomic clone fished out with probes homologous to the AprD sequence which is the ABC domain of a *Pseudomonas aeruginosa* alkaline protease. A PSI-BLAST search of the BpABC1 protein sequence reported significantly relevant hits to known or probable nucleotide binding domains of bacterial ABC transporters for the first 20 non-redundant sequences. An extended survey for the following non-redundant BLAST hits (50 were surveyed) showed that the archived database sequences all belonged to proteins from an ABC transport pathway.

From the PSI-BLAST search, the closest alignments which identified the substrate transported were for sugars and branched-chain amino acids. We therefore, believe that the predicted sequence represents the ATP binding subunit of an ABC transport pathway with similar function. The twenty closest homologues from non-redundant species of the BLAST results were retrieved from GenBank for multiple sequence alignment. The sequence of a *S. typhimurium* histidine permease extracted from its crystal structure PDB file was also included in the alignment to provide an integrated view of the sequence analysis and structure prediction research (Table 1). Analysis of the ClustalW results revealed that the aligned sequences showed three clear and consistent conserved regions (Fig. 1). These regions of conservation were identified and considered so because they contained at least two residues which were conserved throughout the alignment. The conserved regions were further identified as the region containing the Walker A/B nucleotide binding motif as well as a glycine rich motif with the majority having a motif of LSGGQ. We found that these 22 sequences while clearly homologous were considerably variable in other parts of the primary protein structure. A further comparison by BLASTN at nucleotide level yielded no significant hits to each other as expected for this level of protein diversity.

Table 1: Identification labels and definitions for sequences used in comparative analysis. GenBank protein database accession numbers were used for all sequences except for 1B0U which utilized a PDB ID. The protein accession number AAK67358 is synonymously used with the nucleotide accession number AF390557 with the nucleotide entry being the data submitted to GenBank

Accession	Organism	Length	GenBank definition
Burkholderia pseudomallei			
AF390557		264	Probable ABC transporter
NP_360137	Rickettsia conorii	240	Putative ABC transporter ATP binding protein
NP_220752	Rickettsia prowazekii	247	ABC transporter ATP binding protein
NP_358751	Streptococcus pneumoniae R6	180	Conserved hypothetical protein
BAB62754	Streptococcus criceti	252	Putative ABC transporter ATP binding protein
NP_607141	Streptococcus pyogenes	252	Putative ABC transporter ATP binding protein
NP_519859	Ralstonia solanacearum	264	Probable ATP binding protein
NP_252527	Pseudomonas aeruginosa	264	Probable ATP binding protein
NP_230748	Vibrio cholerae	264	ABC transporter, ATP binding protein
NP_355607	Agrobacterium tumefaciens	264	AGR_C_4841p
NP_387317	Sinorhizobium meliloti	264	Putative ATP binding protein, ABC transporter
NP_538930	Brucella melitensis	264	ABC transporter, ATP binding protein
NP_601448		263	COG1101:Various ABC transport systems, ATPase components
Corynebacterium glutamicum			
AAF86640	Enterococcus gallinarum	269	ABC transporter
AAL73232	Streptococcus gordonii	267	ABC transporter
NP_102392	Mesorhizobium loti	259	probable ATP-binding component of ABC transporter
NP_602864	Fusobacterium nucleatum	258	ABC transporter ATP binding protein
NP_563289	Clostridium perfringens	285	probable ABC transporter
NP_349701	Clostridium acetobutylicum	286	ABC-type transporter, ATPase component (cobalt transporters subfamily)
NP_107546	Mesorhizobium loti	249	ABC transporter, ATP binding protein
1B0U_A	Salmonella typhimurium	262	Chain A, ATP binding subunit of the histidine permease
NP_396606	Agrobacterium tumefaciens C58	271	AGR_pTi_165p

The significant overall alignment achieved serves as a secondary level of confirmation for the accuracy of our gene prediction for BpABC1. The alignments also revealed a glycine residue (Gly61) and a leucine residue (Leu179) which was invariably conserved across all sequences and was not associated with any of the three motifs discussed.

The multiple sequence alignment data was used to create a phylogenetic tree of the same non-redundant species as used in the alignments (Fig. 2). This analysis confirmed the diversity of the profile built from the multiple sequence alignment.

The BpABC1 sequence was found to be most closely related a probable ATP binding protein from *Ralstonia solanacearum*. The next closest relationship was shown towards a probable ATP binding protein from *P. aeruginosa*.

These findings corroborate known reports of taxonomic studies for the *Burkholderia*, *Pseudomonas* and *Ralstonia* genuses. *Pseudomonas solanacearum* and *Burkholderia solanacearum* are synonymous with *Ralstonia solanacearum*. *B. pseudomallei* was also formerly classified as a *Pseudomonas*.

The evolutionary relationships helped to create a distinction between the different species compared but does not serve as an accurate analysis of phylogenetic relationships. This is mainly due to the fact that the (substrate) substance transported may belong to different pathways and therefore resulting in the proteins compared also belonging to different families in addition to the phylogenetic differences. The multiple sequence alignment and phylogenetic tree therefore, reveals information of either phylogenetic differences or evolutionary relationships or a combination of both. Due to the objective of refining gene prediction accuracy via these tools, we found that these observations while important to note did not interfere and to an extent corroborated our gene identification to a better biological level of confidence.

Nevertheless, it would be worth to note that the alignments and phylogenetic tree also revealed a possibility for the existence of three different families of ABC transporters in the analysis (Fig. 1). These clusters may in turn be dependent on the type of substance transported.

The inclusion of a wild card sequence in 1B0U also revealed that it integrated well with the rest of the alignment and was not isolated. This shows the high degree evolutionary conservation for the three crucial regions of the NBDs aligned.

We observe this based on the fact the 1B0U entry was distant in the HSPs list for BLASTP and was yet able to integrate into the alignment despite the distance. Again, the use of the term distant reflects on a single or combination of possibilities as discussed above. This integration further demonstrates the evolutionary or phylogenetic conservation required for a primary structure to achieve a specific function (in this case ATP

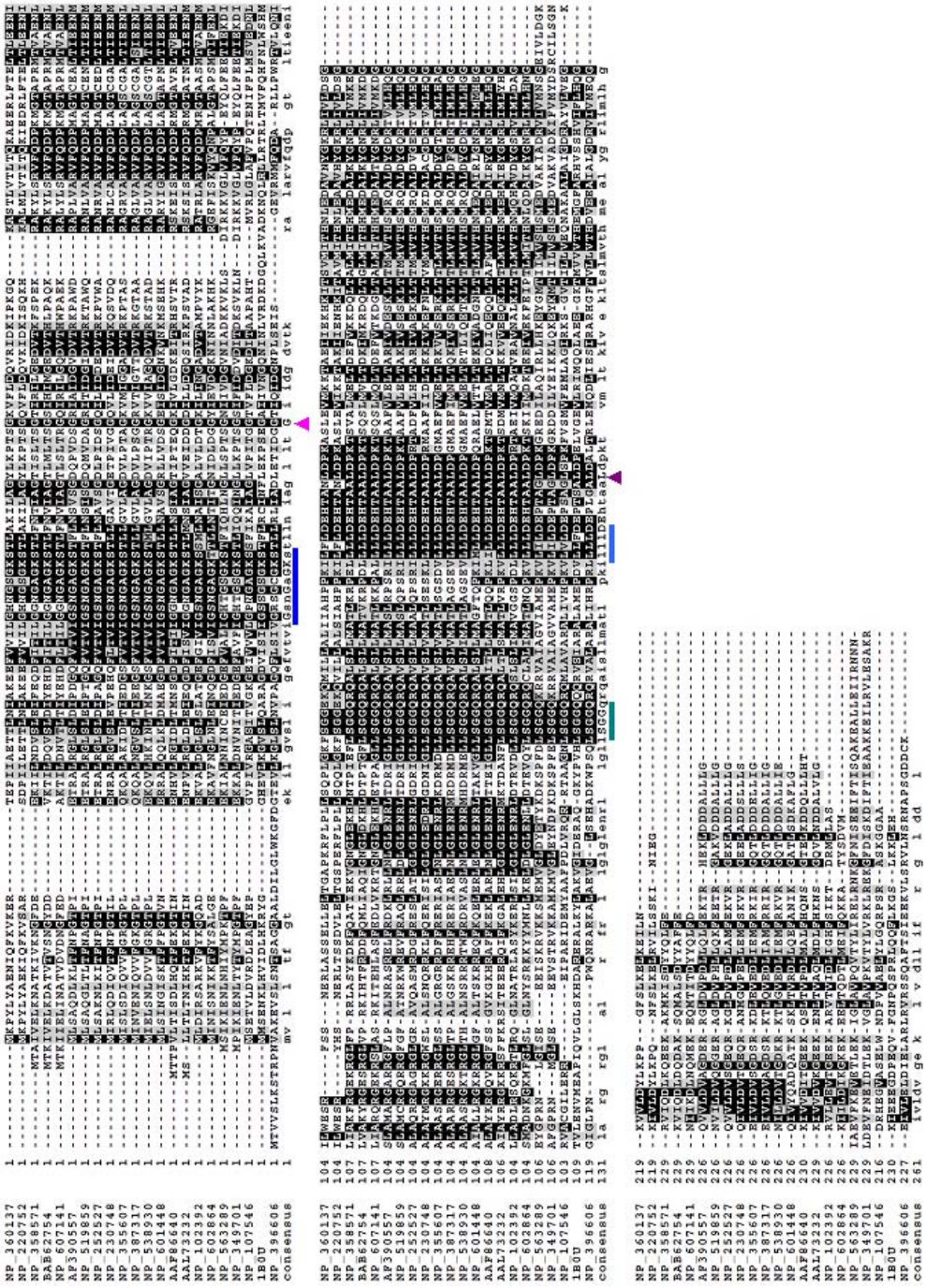

Fig. 1: Multiple sequence alignment of a bacterial ATP binding cassette sequences. GenBank accession numbers and a PDB ID was used to label the alignment sequences with definitions available in Table 1. Invariant residues are highlighted in black and conserved residues are highlighted in gray. The conserved Walker A/B motifs are marked with lines in blue shades while the LSGGQ motif is marked by a green line. The conserved Gly61 and Leu179 residues are marked by triangles in purplish shades

binding). This information was used as a basis for inference of function based on tertiary structure and known mechanisms of ATP binding based on the primary structure information.

Structure prediction and comparative modeling: The results of template searching yielded the histidine permease crystal structure (PDBID = 1B0U) from *S. typhimurium* as the most suitable template structure

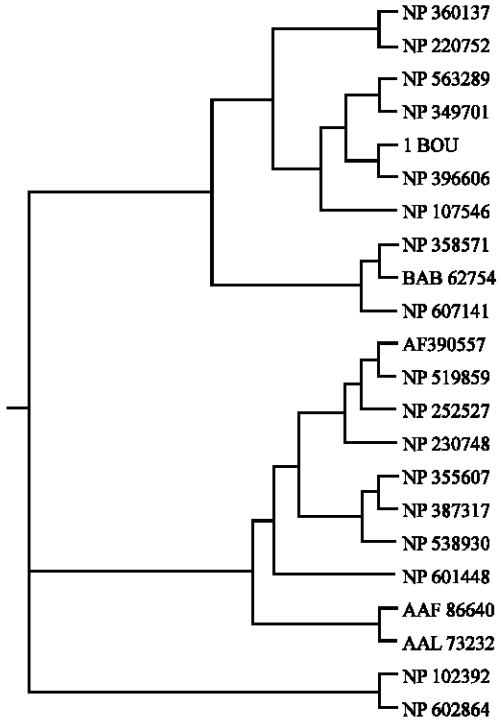

Fig. 2: Phylogenetic tree of a bacterial ATP binding cassette family GenBank accession numbers and a PDB ID was used to label the tree with definitions available from Table 1

for comparative modeling. Transmembrane prediction tools utilized detected no significant transmembrane regions. This step while not crucial, served to discount the possibility of the sequence being a transmembrane component of the ABC transporter proteins. The results of this analysis had already been pre-confirmed by the multiple alignments containing the conserved NBDs. The reason for still carrying out, this was due to the fact of many GenBank entries carrying sequences for the whole pathway inclusive of the transmembrane and outer-membrane component of the system.

An optimal pairwise alignment of the BpABC1 sequence to 1B0U revealed a 26.4% identity (Fig. 3). Secondary structure prediction utilizing the Kabsch and Sander (1983) secondary structure definitions were done on both the target and template sequences for comparisons. The secondary structures from the solved crystal structure was used to gauge structural similarity in the target BpABC1 sequence (Fig. 3). The secondary structure comparison revealed that the Walker A motif was located on a loop. This observation corroborates functional details of the Walker A motif being a phosphate binding loop or P-loop as proven by Hung *et al.* (1998).

Multiple alignments of the twenty best non-redundant results from the PSI-BLAST search along with the template structure sequence shows the conserved ATP binding motifs of Walker A and B and a predominantly LSGGQ motif which have been reported as highly conserved in the NBDs of ABC transporters (Fig. 1) (Venclovas *et al.*, 2001). The glycine rich region containing the LSGGQ motif has been reported as a hot spot for mutations in cystic fibrosis transmembrane conductance regulator proteins, an ABC transporter protein of importance in cystic fibrosis clinical manifestations (Cutting *et al.*, 1990).

The model of the ABC protein shows a hydrophobic core and a mostly hydrophilic surface which was in agreement with the Kyte and Doolittle hydropathy plot. The two conserved NBD motifs are in a cytosol exposed hydrophilic cleft. These observations correspond to the globular and cytoplasmic nature of the ABC transporters which carry the NBD. The model was biologically significant when compared with corroborating evidence of known structures and mechanisms.

We have further assessed the quality of the model with the programs Procheck (Laskoswki *et al.*, 1993), ERRAT (Colovos and Yeates, 1993) (Fig. 4) and Verify3D (Luthy *et al.*, 1992). All residues have also been determined to be in the allowed regions of the Ramachandran plot (Ramachandran *et al.*, 1968). Regions of bad geometry were mainly due to poor sequence alignments.

Function inference by predicted structure and extrapolative mechanisms comparisons: The model we have predicted consists of 7 clear alpha helices and β-11 sheets. We also observed, a Rossman fold for the model which consists of five parallel central beta sheets flanked by helices (Fig. 5a, b). The Rossman fold is known to be synonymous with structural domains that can efficiently bind ATP (nucleotide binding domains) and facilitate its hydrolysis (Rao and Rossmann, 1973). The Walker A and B nucleotide binding motifs were observed to be within this region for the model as well as the template structure (Fig. 5a, b). This fold may be conserved structurally for other NBDs within this protein super family. It is therefore, believed that other future models or structures of the ATP binding cassette may share this topology due to its apparent contribution to what may be the proper folding and conformation necessary for ATP binding and ATPase activity.

The overall structure of the histidine permease (HisP) template is shaped like an L with two thick arms (arm I and II); this monomer forms a dimer which is related by a 2 fold axis *in vivo* (Hung *et al.*, 1998). The model as expected,

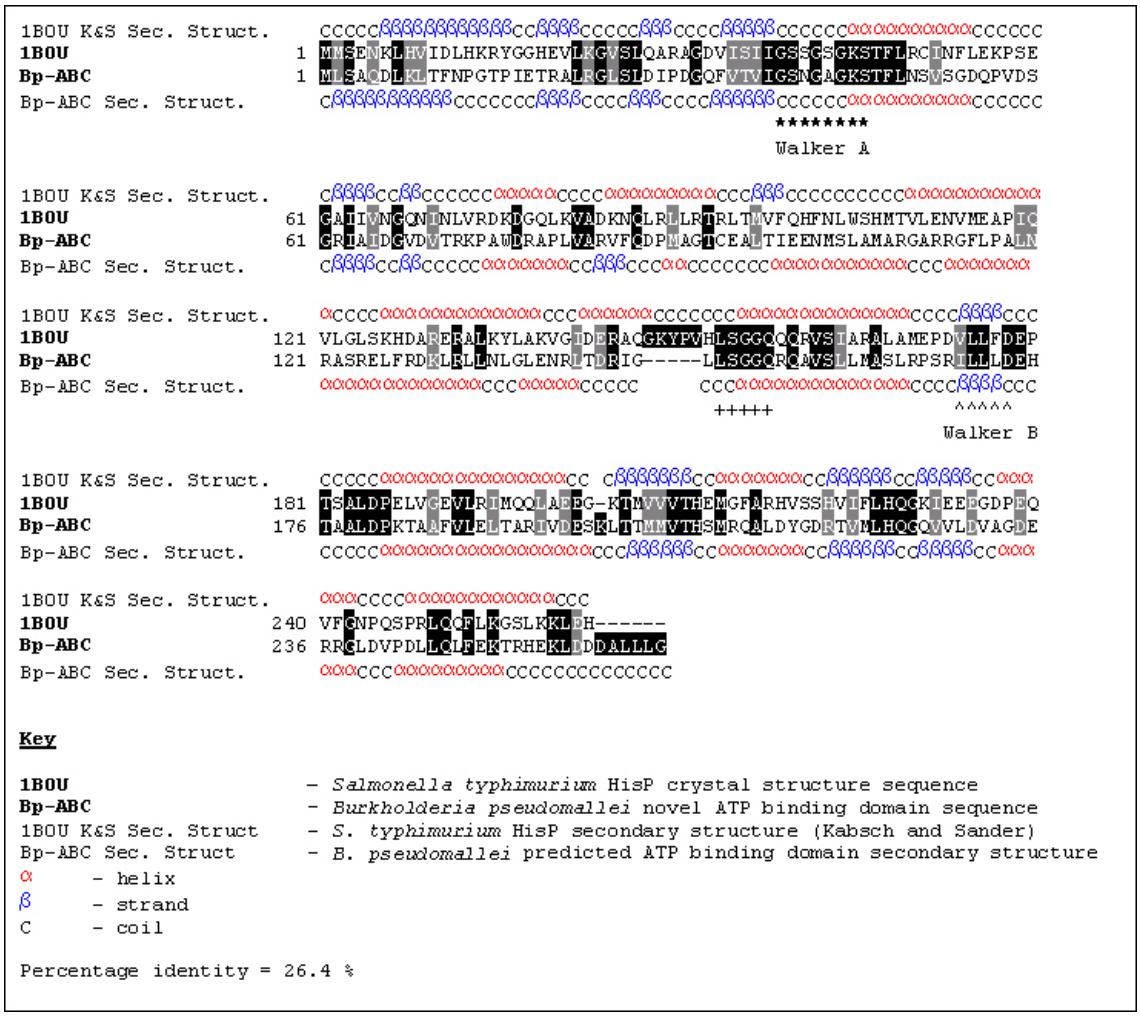

Fig. 3: Pairwise target-template sequence and secondary structure alignment. Identical residues are highlighted. Kabsch and Sander definition of secondary structures was used and the secondary structures of the target and template was compared

shares a similar overall structure with the template. The Walker A motif is located at the lower end of arm I on a glycine rich loop which joins the β-6 strand to the α-1 helix. The Walker B motif was found located near the top end of arm I on the β-11 strand. Both the β-6 and 11 strands are a part of the five central parallel beta sheets which form a part of the Rossman fold.

Another key observation that strengthens the plausibility of the model is the close position to each other in the structure of the two Walker motifs (Fig. 5a, b) resulting in the formation of a pocket like indentation which is known to be involved in ATP binding and hydrolysis. This is believed to be an important factor to gauge proper theoretical folding due to the considerable distance of the two consensus sequences in the primary structure as compared to the closeness of those motifs in the tertiary structure model. The Asp173 residue of the Walker B motif is proximally the closest to the Lys45 and Ser46 residues of the Walker A motif. We have also identified, the region that may most likely be the ATP binding domain due to it containing both the Walker A and Walker B motifs (Fig. 5c). This particular region has already been confirmed on the template structure used (Hung *et al*., 1998). The ATP binding pocket contains the phosphate-binding loop and includes residues Gly39 to Lys45 (Hung *et al*., 1998). The LSGGQ motif was however located on a loop at the end of arm II (Fig. 5c) and not in the region of the two Walker motifs. Even though this motif is located near the surface of the protein, it is believed that it does not directly play a role in ATP binding and hydrolysis. This glutamine rich sequence has been termed as a linker peptide.

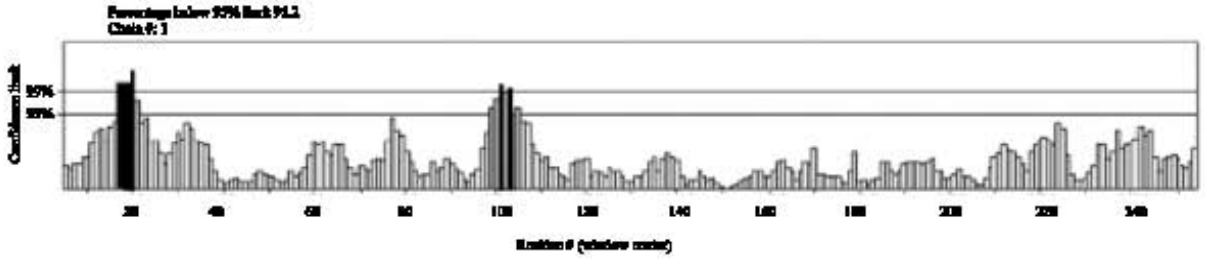

Fig. 4: Results of predicted tertiary structure evaluation by Errat. The Errat program evaluated a limit of 95.2% below the 95% confidence limit; a minimum of 95% is required to be considered a good structure

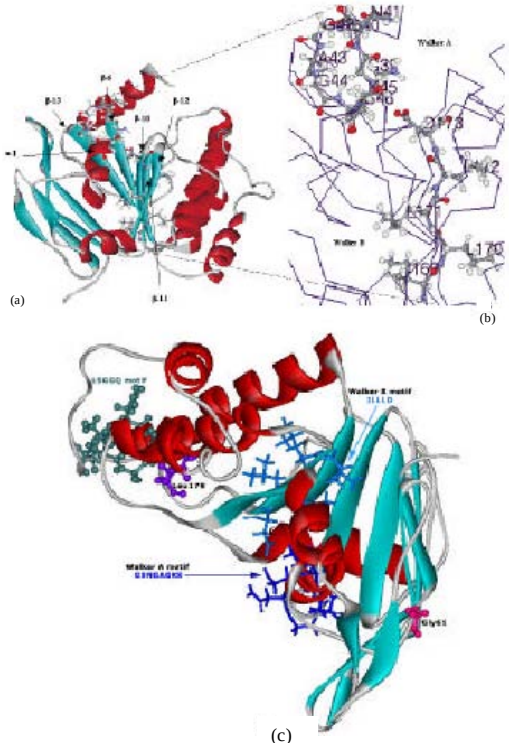

Fig. 5: Predicted tertiary structure model of the BpABC1 protein. a) Diagram of the predicted model showing the Rossman fold; b) Magnification of the diagram a with the backbone rendered as sticks; c) Predicted model of BpABC1 in the orientation to form a dimer with functionally important motifs and residues discussed in the text rendered and labeled for differentiation

Despite the existence of a similar solved structure (HisP), the true function of this highly conserved motif remains uncertain. Hung believed that this motif may be essential for the integrity of the folded protein in the case of HisP (Hung *et al*., 1998). We would also like to present the possibility of it being a crucial region for interacting with the other proteins in the ABC transport pathway. These interactions may be indirectly involved in ATPase activity by ensuring the correct structural conformation to enable ATP binding and ATPase activity. There is a further possibility the motif may play a particular role in interacting with the integral membrane proteins. This is due to the reason that this particular motif, though hydrophilic is surrounded by a hydrophobic region which may possibly interact with the inner membrane. Mutations to this fragile region may then result in a disability to properly interact with the inner membrane due to conformational changes and thereby resulting in the loss of ATP driven transmembrane transport. The conserved Gly61 and Leu179 residues detected in the alignments

were found to be placed in potentially crucial locations in the tertiary structure. From the primary protein structure, these residues appear to be independent of the three known motifs, Walker A/B and LSGGQ. The multiple alignments done did not yield sufficient information on the possible biological functions of these residues.

Analysis of the predicted tertiary structure enabled us to present hypotheses that these residues may play important roles in structural integrity maintenance of the ATP binding subunit structure. The Gly61 residue was observed in a position near the surface of the protein facing the other monomer of the ATP binding dimer complex (Fig. 5c). Therefore, this residue may have roles in interactions which are crucial for the integrity of the assembled ABC transporter complex. Another possibility observed was the role that the Gly61 residue may play in maintenance of the general L shaped curvature of the monomer and thus, contributing to the proper formation of the ATP binding pocket.

Should this be the case, the Gly61 residue may in effect be an extended part of the Walker A motif for family members closely related to the BpABC1 protein. This possible motif extension, even though detectable in the primary structure seemed only functionally explainable with the assistance of tertiary structure data. The Leu179 residue was observed to be close to the LSGGQ motif in the tertiary structure (Fig. 5c). In the predicted NBD, the Leu179 is positioned between the LSGGQ and Walker B motif in structure space. It could therefore also, affect activity in a way similar to the influence of the LSGGQ motif. It is also probable that Leu179 is required as a bridging factor.

Mutations to either this residue, the LSGGQ motif or the Walker B motif may result in improper formation of the ATP binding pocket. Hence, mutations in LSGGQ may also affect the ATP binding pocket. We are unable to present conclusive explanations due to the lack of experimental data discussing the sequence, structure and function relationships for these residues. We note that the Gly61 and Leu179 conservation may be limited to a specific family or sub-family of ABC transporters only and not conserved throughout the super family as such.

CONCLUSION

High resolution experimentally solved structural models of bacterial ABC transporters are still very rare. A survey of the CATH (Orengo *et al*., 1997) SCOP (Murzin *et al*., 1995) and FSSP (Holm and Sander, 1996) databases for structural neighbours of the HisP template yielded no other proteins with similar fold. We demonstrated a high biological confidence in the prediction of the gene and corresponding protein structure of a biologically uncharacterised *B. pseudomallei* ATP binding protein from the ABC transporter super family. Even though, the NBDs of ABC transporters may have wide ranging levels of sequence identities, motifs pertaining to the ATP binding site as well as a few other motifs with yet to be fully explained functions have been shown to be highly conserved in the primary sequence and also structurally. The model has proven that these characteristics could therefore be taken advantage of to predict the structures and infer function of other ABC proteins regardless of the substrate being transported in that particular pathway.

These predicted structures as shown can successfully and satisfactorily explain the structure from a functional point of view despite the apparent lack of suitable highly homologous templates. This will enable the quick yet refined study of other ABC transporter mediated pathways with pathological significance such as for toxins, proteases, drug efflux as well as for diseases like cystic fibrosis via computational analysis using the limited solved structures available for functional determination of probable ABC proteins.

ACKNOWLEDGEMENTS

This research was funded by the grants IRPA 09-02-02-T001 and UKM-MGI-NBD0002-2007 from the Ministry of Science, Technology and Innovation, Malaysia.

REFERENCES

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res., 25: 3389-3402.

Berman, M.H., J. Westbrook, J. Feng, G. Gillilang and T.N. Bhat *et al*., 2000. The protein data bank. Nucl. Acids Res., 28: 235-242.

Brett, P.J. and D.E. Woods, 2000. Pathogenesis of and immunity to melioidosis. Acta Trop., 74: 201-210.

Colovos, C. and T.O. Yeates, 1993. Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Sci., 2: 1511-1519.

Cserzo, M., E. Wallin, I. Simon, G. von Heijne and A. Elofsson, 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: The dense alignment surface method. Protein Eng., 10: 673-676.

Cutting, G.R., L.M. Kasch, B.J. Rosenstein, J. Zielenski, L.C. Tsui, S.E. Antonarakis and H.H. Kazian, 1990. A cluster of cystic fibrosis mutations in the first nucleotide-binding fold of the cystic fibrosis conductance regulator protein. Nature, 346: 366-368.

Dance, D.A.B., 2002. Melioidosis. Curr. Opin. Infect. Dis., 15: 127-132.

Fischer, D. and D. Eisenberg, 1996. Protein fold recognition using sequence-derived predictions. Protein Sci., 5: 947-955.

Fischer, D., 2000. Hybrid fold recognition: Combining sequence derived properties with evolutionary information. Pac. Symp. Biocomput., 2000: 119-130.

Hase, C.C. and R.A. Finkelstein, 1993. Bacterial extracellular zinc containing metalloproteases. Microbiol. Rev., 57: 823-837.

Henikoff, S., E.A. Greene, S. Pietrokovski, P. Bork, T.K. Attwood and L.G. Hood, 1997. Gene families: The taxonomy of protein paralogs and chimeras. Science, 278: 609-614.

Holm, L. and C. Sander, 1996. Mapping the protein universe. Science, 273: 595-602.

Hung, L.W., I.X.Y. Wang, K. Nikkaido, P.Q. Liu, G.F.L. Ames and S.H. Kim, 1998. Crystal structure of the ATP-binding subunit of an ABC transporter. Nature, 396: 703-707.

Jones, D.T., 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol., 287: 797-815.

Jones, D.T., W.R. Taylor and J.M. Thornton, 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry, 33: 3038-3049.

Jr. Brosh, R.M. and S.W. Matson, 1995. Mutations in motif II of *Escherichia coli* DNA helicase II render the enzyme non-functional in both mismatch repair and excision repair with differential effects on the unwinding reaction. J. Bact., 177: 5612-5621.

Kabsch, W. and C. Sander, 1983. How good are predictions of protein secondary structure. FEBS Lett., 155: 179-182.

Karplus, K., C. Barrett, M. Cline, M. Diekhans, L. Grate and R. Hughey, 1999. Predicting protein structure using only sequence information. Proteins, 3: 121-125.

Kyte, J. and R.F. Doolittle, 1982. A simple method for displaying the hydropathic character of a protien. J. Mol. Biol., 157: 105-132.

Laskoswki, R.A., M.W. MacArthur, D.S. Moss and J.M. Thornton, 1993. PROCHECK: A program to check the stereochemical quality of protein structures. J. Applied Cryst., 26: 283-291.

Lim, A. and L. Zhang, 1999. WebPHYLIP: A web interface to PHYLIP. Bioinformatics, 15: 1068-1069.

Luthy, R., J.U. Bowie and D. Eisenberg, 1992. Assessment of protein models with three-dimensional profiles. Nature, 356: 83-85.

Matte, A., H. Goldie, R.M. Sweet and R.M. Delbaere, 1996. Crystal structure of *Esherichia coli* phosphoenol-pyruvate carboxykinase: A new structural family with the P-loop nucleoside triphospahte hydrolase fold. J. Mol. Biol., 256: 126-143.

Murzin A.G., S.E. Brenner, T. Hubbard and C. Chothia, 1995. Scop A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., 247: 536-540.

Orengo, C.A., A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells and J.M. Thornton, 1997. CATH-a hierarchic classification of protein domain structures. Structure, 5: 1093-1108.

Putman, M., H.W. van Veen and W.N. Konings, 2000. Molecular properties of bacterial multidrug transporters. Microbiol. Mol. Biol. Rev., 64: 672-693.

Ramachandran, G.N., C. Ramakrishnan and V. Sasiskharan, 1968. Stereochemistry of polypeptide chain configurations. Adv. Protein Chem., 23: 283-437.

Rao, S.T. and R.M.G. Rossmann, 1973. Comparison of super-secondary structures in proteins. J. Mol. Biol., 286: 241-256.

Sali, A. and T.L. Blundell, 1993. Comparative protein modelling by the satisfaction of spatial restraints. J. Mol. Biol., 234: 779-815.

Shi, J., T.L. Blundell and K. Mizuguchi, 2001. FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol., 310: 243-257.

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: 4673-4680.

Venclovas, C., A. Zemla, K. Fidelis and J. Moult, 2001. Comparison of performance in successive CASP experiments. Proteins, 5: 163-170.

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

Wandersman, C., 1998. Protein and peptide secretion by ABC exporters. Res. Microbiol., 149: 163-170.