Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry
Computational Biology and Chemistry 31 (2007) 115–123
www.elsevier.com/locate/compbiolchem

Ligand-binding prediction in the resistance-nodulation-cell division (RND) proteins

Armando Hernandez-Mendoza [a], Carmen Quinto [a], Lorenzo Segovia [b], Ernesto Perez-Rueda [a,*]

[a] *Departmento de Biología Molecular de Plantas, Instituto de Biotecnologia, Universidad Nacional Autónoma de México, A.P. 565-A Cuernavaca, Morelos 62210, Mexico*
[b] *Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnologia, Universidad Nacional Autónoma de México, A.P. 565-A Cuernavaca, Morelos 62210, Mexico*

Received 18 May 2006; received in revised form 7 February 2007; accepted 8 February 2007

Abstract

The resistance-nodulation-cell division (RND) protein family is a ubiquitous group of proteins primarily present in bacteria. These proteins, involved in the transport of multiple drugs across the cell envelope in bacteria, exhibit broad substrate specificity and act like efflux pumps. In this work, a protein belonging to the RND protein family, AcrB of *Escherichia coli* was used as a working model to predict *in silico* the compounds transported by 47 RND proteins. From AcrB we extracted and clustered 14 amino acids directly involved in substrate interactions. This clustering provides enough information to identify 16 groups that correlates with the ligand they extrude, such as proteins expelling aromatic hydrocarbons (SrpB cluster) or proteins expelling heavy metals (CnrA cluster). The relationship between conserved, cluster-specific and variable residues indicates that although the ligand-binding domain is conserved in structure, it has enough flexibility to recognize specifically a diversity of molecules.

Keywords: RND-proteins; AcrB; Ligand binding prediction

1. Introduction

In general, homologous proteins share common biological functions but exhibit different specificity towards their substrates. The identification of residues that participate in the specificity of these interactions is useful for functional analysis, and for predictive studies. Recent analyses have shown that residues involved in specific recognition between interacting partners are frequently found at conserved positions (Johnson and Church, 2000). These positions are evident in multiple sequence alignments, where the distribution of amino acids reflects functional and structural constraints in the proteins. Recently, a number of algorithms addressing ligand-binding predictions have been developed, such as the Evolutionary Trace method, that exploits information on protein sequence and structure (Lichtarge et al., 1996), whereas alternative methods only use protein sequence information (Berezin et al., 2004).

The resistance-nodulation-cell division (RND) protein family is a ubiquitous group of proteins described in bacteria, archaea and eukarya, which is involved in the transport of multiple drugs across the cell envelope (Paulsen et al., 1996; Paulsen et al., 1997). The RND proteins form complexes with membrane fusion proteins (MFP) in the periplasm and with outer membrane channels of TolC superfamily to accomplish the transport from the cell to the extracellular medium. Diverse RND efflux systems have been functionally characterized, which are involved in the transport of and resistance to antibiotics, hydrophobic dyes, and detergents, among others (Poole, 2004). In addition, a large number of point mutations causing altered substrate or inhibitor specificities in different multidrug transporters have been reported (Hearn et al., 2006; Murakami et al., 2004; Yu et al., 2005) as well as crystallographic data of purified AcrB in the presence of several ligands (Yu et al., 2003). These results together do not only suggest that these proteins form direct atomic interactions with the molecules of substrates but a dif-

* Corresponding author. Tel.: +52 56 22 76 10; fax: +52 777 3 17 23 88.
E-mail address: erueda@ibt.unam.mx (E. Perez-Rueda).

1476-9271/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2007.02.003

ferential interaction can be achieved as a consequence of the substrate nature.

In order to predict ligand molecules in the RND protein family in bacteria, we clustered 14 columns that correspond to amino acids selected from the AcrB protein of *Escherichia coli*, whose 3D structure has been determined in the presence of four structurally diverse ligands (Murakami et al., 2002). *E. coli* AcrB is an excellent archetype to predict ligand binding molecules, since it recognizes many structurally unrelated toxic compounds, such as antibiotics, disinfectants, dyes, and simple solvents (Nikaido and Zgurskaya, 2001). In this work, a multiple sequence alignment (MSA) was constructed with 47 RND-like protein sequences and 14 columns were extracted corresponding to residues directly involved in substrate interactions into the central cavity (Yu et al., 2003) of AcrB and to residues located in the central pore (Murakami et al., 2002). These 14 columns were used to generate a tree that shows diverse RND multidrug efflux family clusters that correlate with the ligand they recognize and extrude across the membrane cell. Similar approaches have shown the advantage of anchoring functional annotations within a protein family context (Johnson and Church, 2000).

2. Methods

2.1. Sequence analysis

The *E. coli* AcrB protein sequence, whose crystallographic structure was solved at 3.5 Å, was used as seed in a Blast search with an *E*-value $\leq 10^{-5}$ to identify RND protein members in the NR-database. Five hundred and sixty-five protein sequences were retrieved from this sequence comparison, but only 43 RND-proteins were selected whose ligands are well known. In addition, RmiB and ORF2pC (*Rhizobium etli*), Orf2 (*R. leguminosarum*) and AmeB (*Agrobacterium tumefaciens*) were included in the analysis because we have experimental evidences supporting a different function than previously reported. See below.

2.2. Multiple sequence alignments (MSA) and ligand-binding-site analysis

Forty-seven RND proteins were aligned using ClustalX with default conditions and manually edited (Thompson et al., 1997, 1994). In this alignment the AcrB protein sequence from *E. coli* K12 was used as reference to identify and extract 10 amino acid residues involved in ligand binding. Those residues are 6 Å or less from their ligands (forming preferentially hydrophobic, Van der Waals, or electrostatic interactions) in the AcrB protein structure (1IWG) (Yu et al., 2003). Four additional amino acid residues identified experimentally by mutational analysis were also included (Murakami et al., 2004). In total, 14 amino acid residues were selected to perform the analysis: five residues are in the central pore (D101, V105, N109, Q112 and P116), and nine in the central cavity (L25, K29, D99, V382, A385, F386, F388, F458 and F459) (see Table A.1, Fig. A.1 in Appendix A).

In a second step, columns from the MSA corresponding to the 14 ligand-binding residues were selected and used to build a new alignment. This new alignment was analyzed by maximum parsimony (MP), fitch (F) and neighbor-joining (NJ) methods, and their corresponding trees were generated. From this analysis, 16 clusters subsequently emerged that correlated with the substrates they extrude. The MP tree clusters were used to predict ligands in those proteins whose ligands are unknown or contradictories. Alternatively, 100 random MSAs with the same amino acid composition and length than the original were constructed to evaluate the previously described clustering. We must emphasize that phylogenetic tools were only used to cluster similar sets of residues, not to make inferences on the evolutionary history of these proteins.

2.3. Performance evaluation

Finally, to evaluate the performance of the ligand-binding predictions, the clusters identified by MP ("observed clusters") were compared to their corresponding annotated ligands. This comparison was useful to calculate the following values: (1) true positives (TP): proteins with (at least one) common ligand clustered together; (2) false positives (FP): proteins whose ligand was completely different to the rest of the cluster; (3) false negatives (FN): proteins included in a cluster with different ligands; (4) sensitivity, Sn = TP/(TP + FN), is the fraction of proteins recovered in the inferred clustering; (5) positive predictive value, PPV = TP/(TP + FP), is the fraction of the proteins and ligands in the inferred clusters that belongs to the annotated ligand binding; (6) Accuracy, Ac = (Sn + PPV)/2, is the PPV and Sn average.

In this analysis the classical definition of specificity Sp = TN/(TN + FP) was not used, because our evaluation criteria was based on the rate of true negatives (TN), defined as proteins whose ligand has not been experimentally described. Indeed, the number of ligands is typically smaller than the size of the cluster (16 clusters), and the percentage of TN should be always closer to 1, which would favorably bias in the evaluation.

3. Results

3.1. Identification and selection of RND proteins

In order to identify members of the RND family, the *E. coli* AcrB protein sequence was used as seed to scan the NR database. Five hundred and sixty-five proteins were identified as potential RND-proteins and 43 proteins functionally characterized in diverse bacteria were selected, ranging from proteobacteria to bacteroidetes (Fig. 1). These proteins share in average 40% of identity among them, and the information on their ligand molecules is known. Around 90% of the 565 proteins identified by BLAST search have been annotated as hypothetical, unknown or uncharacterized proteins, emphasizing the importance of predict their probable ligand compound. In addition, four proteins RmiB and ORF2pC (*R. etli*), Orf2 (*R. leguminosarum*) and AmeB (*A. tumefaciens*) whose ligands are unknown or contradictories were considered in this analysis. AmeB might be involved in ß-lactams and detergents extrusion, however its phenotype is not completely clear when the protein is deleted (Peng