Chapter 2

Bioinformatics analysis of the ECF transporter subunits

Summary

This chapter contains a bioinformatics analysis of all ECF transporter subunits. First, a phylogenetic study on Nucleotide Binding Domain (NBD) sequences is used to determine the relation between ECF- and ABC transporters. Subsequently, conserved sequence motifs and the predicted topology of the EcfT subunit are discussed. Finally, based on multiple sequence alignments of S-component sequences from *L. lactis*, short sequence motifs are identified that are shared between different S-components. It is proposed that these motifs are structurally important and indicative of a general S-component fold.
Introduction

Research on ECF transporters is still in its early days and our insight in the mechanism of transport is therefore limited. Nevertheless, ECF transporters genes are identified in many prokaryotes (108) and their sequences are a good starting point for bioinformatics analysis. In addition, there is a clear link between the well-studied ABC transporter family and ECF transporters, through the conserved Nucleotide Binding Domains (NBDs) that energize the transport reaction. To understand the exact relation between ABC- and ECF-transporters the sequences of the NBDs can be compared. The most recent phylogenetic analysis of NBD sequences was carried out more than ten years ago (115) and did not include ECF transporter NBDs. In this chapter, a large-scale phylogenetic analysis of NBD sequences from ECF- and ABC transporters is presented that places ECF transporters in the ABC transporter superfamily. Although the molecular identity of ECF transporters was discovered only three years ago, already two crystal structures of S-components are now available: RibU (155) for riboflavin and ThiT for thiamin (chapter 5). These proteins are unrelated in sequence but have a similar fold. In both structures, conserved amino acids have been identified that are important for substrate interaction and others that may have a structural role (helix packing/kinking). Using the knowledge of the 3D structures, it is now possible to search for similar patterns in S-components that do not share significant sequence similarity with either ThiT or RibU. Such an analysis is performed in this chapter. Based on the results predictions are made about the general structural features of S-components.

Results and discussion

Phylogenetic analysis of ECF- and ABC-transporter NBDs

An evolutionary tree based on the alignment of the 350 amino acids ‘core sequence’ of NBDs is depicted in figure 1. A tree based on an alignment length of a smaller stretch of 200 amino acids, showed essentially the same distribution. The tree is annotated based on the description of the sequences in the conserved domain database. As observed before (27), ABC transporters tend to cluster based on their substrate specificity. For instance, transporters for amino acids (methionine, histidine, glutamine and arginine) form a defined group as well as transporters for iron- or cobalt siderophores. Although the division in functional clusters seems evident, the bootstrap scores at the points of branching are very low. Low scores are usually an indication that the observed patterns have to be analyzed with caution. Low bootstrap scores are more often observed for large datasets (115). The poor scoring does not necessarily imply unreliable branching, but indicates that not all members can be assigned to a particular group with high confidence.
Figure 1: an evolutionary tree of NBDs from ECF and ABC transporters. The tree is created using the PHYLIP package (43) and visualized with the program Dendroscope.
The most recent phylogenetic analysis of NBD sequences was carried out over 10 years ago (115). About 200 sequences were aligned to construct an evolutionary tree. Based on this analysis, an early segregation between ABC importers and exporters was proposed. Such an early division between importers and exporters is neither supported nor disproved by the data presented in this chapter. The branching order of the different groups of importers and exporters could not be determined. In general however, exporters from prokaryotic and eukaryotic origin are more related to each other than importers from bacteria. For instance, the bacterial lipid exporter MsbA is grouped among eukaryotic multidrug exporters from the ABC-C family. An exception is the group of nickel/peptide transporters (bootstrap score=25), which contains both oligopeptide importers as well as peptide exporters. These results might suggest that an inversion of the transport direction has occurred multiple times during evolution, some of which could have occurred before the separation between prokaryotes and eukaryotes.

All NBD sequences from ECF transporters fall in one cluster (figure 1), although the bootstrap score is low at the point of branching. If the depicted clustering is reliable, ECF transporters are more related to each other than to any other ABC transporter subfamily, in spite of the large variation of substrates that is transported by ECF transporters. The results presented in this chapter do not indicate that the ECF transporters NBDs have a special position in the ABC transporter superfamily; therefore ECF transporters should be regarded as ABC transporters.

**Genetic distribution of ECF transporters**

ECF transporters are abundant in prokaryotes and can be found in the genomes of many organisms. In figure 2, the relative abundance of the different types of ECF transporters is plotted for several prokaryotic phyla. Type II ECF transporters (that share an energizing module) are particularly abundant in Gram-positive bacteria. On average, the number of S-components in these genomes is about 12% of the total number of ABC transporters. The occurrence of type II ECF transporters seems to be restricted to Gram-positive bacteria, thermotogales and archea, type I ECF transporters (dedicated to one S-component), are more widely distributed. The average prokaryotic genome contains 1-2.5% of ECF type I transporters (percentage of total ABC transporters). In spirochaetes, this percentage is with an average of ~8% much higher. As discussed before, the differences between Gram-negative and Gram-positive bacteria might be explained by the lack of a confined periplasm in the latter species (26). Gram-positive bacteria often have their soluble ABC transporter SBPs fused to the TMDs or lipid anchored in the membrane (136), whereas Gram-negative bacteria can express SBP freely in the periplasm. The membrane inserted ECF transporter S-components are therefore very suitable for utilization by Gram-positive bacteria.
Of all the ECF transporter subunits, the T-component (EcfT) is the most elusive. Although it is the hallmark of ECF transporters, its exact function and role in the transport mechanism is unknown. As the second membrane component in an ABC transporter complex, it might be expected to have a somewhat similar structure as the S-component, because so far all ABC transporters were found to have pseudo symmetry in their transmembrane domains. But topology predictions on EcfT sequences seem to suggest a different structural organization than for S-components (95). The predicted number of transmembrane segments ranges from 5-7 for different EcfT homologues. The variation is mostly concentrated in the C-terminal part of the sequences, which is predicted to be a cytoplasmic domain in about 50% of the sequences, and in the other cases it is predicted to form one or two transmembrane helices. Surprisingly, a substantial amount of sequence conservation localizes in this domain. For example, two motifs containing conserved arginines that were found to be important for the stability of the ECF complex are found here (95). The length of this domain is also very constant among the EcfT orthologues (~100 amino acids) suggesting that it might be of functional importance.
In addition to the arginine motifs, a proline followed by an AxxxA motif is well conserved. This arrangement is somewhat similar to sequence conservation patterns observed in S-components (see below).

**Structurally important residues are shared between S-components**

If all S-components have a similar fold that has originated from a single ancestral protein, their sequences have diverged beyond recognition. The highest sequence identity observed between *L. lactis* S-components is 21% for BioY and HmpT (table 1), but in general the identities are much lower. Even though overall sequence similarity is low, several short sequence motifs were found conserved in (almost) all S-components. For example, the S-components found in *L. lactis* all have an AxxxA motif in their first (predicted) transmembrane helix. Based on the ThiT structure, it is proposed that these alanines might be recognized by EcfT and support interaction of the S-components with the energizing module (chapter 5). In addition to the alanine motif, all S-components except QueT were found to have conserved prolines in helix 2 and 3 (figure 3). A similar conserved proline in ThiT and RibU marks the boundary between the L1 loop and helix H2; in addition it initiates one turn of a 3_{10} helix at this point, and thus serves a very specific structural role. In most S-components the proline is followed by a GxxxG or GxxxA motif, with a distance of 10-15 amino acids. In both the ThiT, and the RibU structure, the first glycine in this motif forms the loop between helix 2 and 3, allowing a very sharp turn of the amino acid backbone at this point. As a result, helix 2 and 3 are packed closely together, which explains the conservation of a second small amino acid (i.e. alanine or glycine); a larger side-chain would not fit in the space between helix 2 and 3. The occurrence of these structural motifs in the *L. lactis* S-components is summarized in figure 3. Although the position of the motifs is not exactly the same in all S-components, there are clearly patterns. If these motifs have a similar function in all S-components, a structural similarity would be very plausible.

### Table 1: pairwise sequence identity (%) between the S-components from *L. lactis* MG1363

<table>
<thead>
<tr>
<th></th>
<th>ThiT</th>
<th>RibU</th>
<th>BioY</th>
<th>PanT</th>
<th>HmpT</th>
<th>QueT</th>
<th>NiaX</th>
<th>BioY2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThiT</td>
<td>-</td>
<td>16</td>
<td>14</td>
<td>19</td>
<td>15</td>
<td>10</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>RibU</td>
<td>16</td>
<td>-</td>
<td>14</td>
<td>19</td>
<td>14</td>
<td>18</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>BioY</td>
<td>14</td>
<td>14</td>
<td>-</td>
<td>8</td>
<td>21</td>
<td>19</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>PanT</td>
<td>19</td>
<td>19</td>
<td>8</td>
<td>-</td>
<td>19</td>
<td>14</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>HmpT</td>
<td>15</td>
<td>14</td>
<td>21</td>
<td>19</td>
<td>-</td>
<td>13</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>QueT</td>
<td>10</td>
<td>18</td>
<td>19</td>
<td>14</td>
<td>13</td>
<td>-</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>NiaX</td>
<td>9</td>
<td>14</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>-</td>
<td>14</td>
</tr>
<tr>
<td>BioY2</td>
<td>7</td>
<td>14</td>
<td>20</td>
<td>21</td>
<td>15</td>
<td>11</td>
<td>14</td>
<td>-</td>
</tr>
</tbody>
</table>
It becomes evident from figure 3 that the similarity between the S-components is concentrated in the N-terminal half of the sequences. The second half does not contain shared sequence motifs and displays a large variation in the arrangement of the predicted transmembrane helices. These differences are also reflected in the structures of ThiT and RibU; helices 1, 2 and 3 are structurally very similar, whereas helix 4, 5 and 6 are more different. Since all S-components interact with the same energizing module this interaction is most likely to take place in a structurally similar part of these S-components. Helices 1, 2 and 3 in the ThiT and RibU structure provide such a surface and probably form a platform for docking of the energizing module and could form the site of interaction with the energizing module.

**Figure 3: shared sequence motifs between different S-components**
The length of the protein sequences is indicated by the gray bars. Sequence motifs and (predicted) transmembrane segments are colored as indicated in the legend. The scale bar corresponds to a sequence length of 20 amino acids.
Methods

Phylogenetic analysis of NBD sequences
The wealth of genomics data and the abundance of ABC transporters creates a challenge for the assembly of a representative set of NBD sequences. For example, in a BLAST search with the ECF transporter EcfA subunit from *L. lactis* (CbiO2) the first 5000 sequences displayed an e-value of less than $10^{-27}$ and were thus most likely orthologues. Because the e-value is so low (even for the last sequence of the list) there are probably many more orthologues in the databases that were not included in the first 5000 hits. In order to acquire a representative set of NBDs from several different types of ABC transporters, a method is needed to select those sequences from the vast amount of ABC transporter NBD sequences that are currently available. For this, we performed a conserved domain search with *L. lactis* EcfA as a query sequence using the CD-search tool (86). The CD-search procedure compares the query sequence with position-specific score matrices that are deposited in the Conserved Domain Database (CDD). In this way, a list of conserved domains was generated that were related to *L. lactis* EcfA. For each of these domains (domains classified as ‘provisional’ were ignored because of uncertainty in their annotation) ten representative sequences were downloaded. A total of 1433 sequences were collected in this way. The set includes ABC transporter NBDs from various biological species (eukaryotes as well as prokaryotes) and from a range of different functional ABC transporter groups. The full-length sequences were aligned with ClustalW (19) using default parameters. Because of the large variation in protein length, the alignment had to be truncated to the conserved ‘core sequence’ of the NBDs. We used the percentage of gaps in blocks of 50 amino acids as a criterion. Each block that contained more than 65% gaps was deleted; this resulted in a final alignment length of 350 amino acids. The whole procedure was repeated with a maximum gap percentage of 53% to determine if the observed clustering was biased by additional domains of the NBDs. This yielded an alignment length of 200 amino acids. An evolutionary tree was constructed with programs from the PHYLIP package (43), to test the reliability of the branching, a bootstrap analysis was performed with 100 replicates.

Genetic distribution of ECF transporter genes
The occurrence of ECF transporter genes in prokaryotic genomes has been described before (108). Based on this data, the number of ECF transporters (type I and type II) in these genomes was calculated. Each type II ECF transporter S-component was counted as one transporter. The total number of ABC transporters in these genomes was extracted from the TransportDB (106) and these number were used to calculate the relative occurrence of ECF transporters.
Shared amino acid motifs between different S-components

The sequences of the S-components from the *Lactococcus lactis* MG1363 genome were chosen as a starting point for the analysis, because it has been demonstrated unambiguously that all eight S-components interact with the same energizing module (130). Besides ThiT and RibU, six additional S-component sequences were analyzed (see table 1). These S-components are likely to share structural features for interaction with the energizing module. Based on each S-component sequence, a multiple sequence alignment was constructed with the program FRpred (45). The alignments were manually searched for patterns of conservation, similar to ThiT and RibU (i.e. GxxxG motifs and conserved prolines). For each *L. lactis* S-component, the conserved amino acids were mapped on a consensus topology prediction by TOPCONS (6) or experimental topology (ThiT and RibU). The pairwise sequence identities were calculated based on a multiple sequence alignment with all eight S-components by ClustalW (19).