Predicting transport proteins for microbiomes: Challenges

Munira Alballa¹, Gregory Butler¹,²*

¹ Dept of Computer Science & Software Eng., Concordia University, Montréal, Québec, Canada
² Centre for Structural and Functional Genomics, Concordia University, Montréal, Québec, Canada
* (Corresponding author) gregory.butler@concordia.ca

Abstract

Transporters mediate the movement of compounds across the membranes that separate the cell from its environment, and across inner membranes surrounding cellular compartments. Transport plays a critical role in metabolism, regulation, and signalling and is particularly important for interaction with microbiomes and pathogens. It is estimated that one third of a proteome consists of membrane proteins, and many of these are transport proteins.

Animals and plants co-habit symbiotically with microbial communities, called microbiomes, which affects their health and growth. Scientists use genomics and metagenomics to study the microbiomes and understand the interaction with the host by the exchange of chemical compounds. Our work will show a scientist the list of transporters for an organism, and also for a microbiome, providing information about potential interactions between them.

Objectives

Our work aims to identify and annotate transport proteins in whole proteomes and meta-proteomes. There are a number of objectives:

(1) Classify membrane proteins in the (meta-)proteome into the functional categories transporter/channel, receptor, enzyme, and other; and into the structural types: single-pass Type I, II, III and IV; multi-pass transmembrane; lipid-anchored; GPI-anchored; and peripheral membrane proteins [2].

(2) Predict the substrate class of a transporter, targeting an extensive range of substrates as in the whole genome annotation [3].

(3) Predict the specific substrate of a transporter.

(4) Scale up the tools to whole meta-proteomes using a compute cluster.

Progress

Our work to date [1] addresses Objective (2), and to a lesser extent Objective (3).

We developed a roadmap for predicting transmembrane transport proteins where the traditional amino acid composition information would be combined with evolutionary information as captured by a multiple sequence alignment (MSA), and by positional information [5] about the residues responsible for determining specificity of the transporter. We also realized the importance of the
alignment preserving the TMS positions since the important residue positions seem to be located there. There are a number of such MSA algorithms, including TM-Coffee [3].

In a preliminary study, we developed TranCEP, for Transporter prediction using Compositional, Evolutionary, and Positional information, which adopts the PAAC encoding scheme, the TM-Coffee MSA algorithm, and the TCS algorithm for determining informative positions in the MSA, to build a suite of Support Vector Machine (SVM) classifiers, one for distinguishing between each pair of classes of substrates. TranCEP significantly outperforms the state-of-the-art TrSSP [4].

Challenges

Objective (3) is extremely challenging. Our previous efforts for de novo prediction of specific substrates for sugar transporters in fungi were not successful. Hence, the roadmap was developed. Much will depend on the number of well-characterized examples available, and the availability of 3D structures.

Objective (4) may require a large-scale cluster for computation, improvements to algorithms, and tuning of analysis pipelines. Computing efficiently at the scale of meta-proteomes is essential. Consider an example meta-proteome, say, with 10M proteins, 3M membrane proteins, and 1.5M transporters. If the average processing time is 10 minutes per sequence, then the total computation takes 6.9 days using 10,000 cores at Compute Canada. Reducing the time for non-transporters to 10 seconds, means a total computation time of 27.2 hours (with 10,000 cores).

References


