ANALYSIS AND ROLE OF TRANSLOCATING PROTEINS IN PROTEIN-PROTEIN INTERACTION NETWORKS

PhD thesis

Péter Mendik MD

Molecular Medicine Doctoral School

Semmelweis University

Supervisors: Péter Csermely DSc
Dániel Veres MD PhD

Official reviewers: Orsolya Kapuy PhD
Bence Szalai MD PhD

Head of the Complex Examination Committee:
Edit Buzás MD DSc

Members of the Complex Examination Committee:
Krisztina Ella PhD
Illés Farkas DSc

Budapest, 2022

Introduction

Subcellular organelles are traditionally defined as compartments of the cells divided by membranes, e.g. mitochondria or nucleus. These subcellular organelles provide their own microenvironment and they enable the spatial separation of different subcellular processes. This separation enables that the intracellular proteins can function at given times with different functions inside different organelles. This altered functionality is partly due to the fact, that in different organelles the same protein may have grossly different interacting partners. This change of interactors naturally explains the observed functional changes.

In living cells there's not only a high level of organization but the constant adaptation to outside stimuli and environment also results in subcellular dynamics. Translocating proteins are prominent players in this always changing cellular world. Translocating proteins move between different organelles, and their movement is a major source of information propagation.

Their appearance in certain organelles is able to change the behaviour of the whole cell.

In general, protein translocation is a process which refers to the alteration of a given protein's subcellular localization. However, this phenomenon has no unified definition, and the word 'translocation' may also refer to gene translocation or RNA translocation at the ribosome. We defined protein translocation as a systems biology phenomenon, which refers to the regulated movement of a protein of a given post-translational state between subcellular compartments.

Protein translocation changes the interaction partners and leads to altered function(s) of translocating proteins. There are certain processes (such as co-translational, post-translational delivery-type, cell division induced, downregulation- or passive diffusion-related phenomena) that may change the localization of a protein, but to increase the focus and clarity of our work we did not consider them as translocation.

Epithelial-mesenchymal transition (EMT) is a biological process which is important during early embryonic

development, but it may also happen during cancer progression or tissue fibrosis. During EMT epithelial cells lose their apical-basal polarity and acquire stem cell like traits. The results in a mesenchymal like cell which is more motile. Mesenchymal cells can also go through a mesenchymal-epithelial transition (MET) which is the inverse of EMT. These transitions are triggered by cellular signals, e.g. Transforming growth factor beta (TGF-β) is a potent inducer of EMT. EMT is defined as a diverse palette of cellular states and cells can be found in any intermediary "hybrid" states between the well-defined epithelial and mesenchymal end states. When we are defining EMT we must also take it into consideration that there is no dedicated marker of EMT (e.g.: the loss of E-cadherin in itself), but one must always assess EMT from a complex approach and simultaneously interpret functional, morphological and transcriptional changes. Prior computational studies of EMT often failed to capture this diverse aspect of EMT.

Translocating proteins play an important role in EMT, e.g. the translocation of β-catenin from the plasma

membrane to the nucleus is a well-known regulatory step of EMT. Although we knew some additional translocating proteins that are important in the regulation of EMT, there was no study that approached EMT from the aspect of protein translocations as regulatory elements of EMT.

Previously Steinway et al. (Cancer Res 74,5963) created a Boolean model of the EMT. That model contained 70 nodes and 135 edges and showed the experimentally known simultaneous activation of sonic hedgehog and WNT pathway during TGF-β mediated EMT. The model is based on experimental data and the *in silico* model rightly recreates experimental outcomes. Through some network reduction steps they have managed to significantly reduce the size of their network (to 19 nodes 70 edges) and this enables the analysis of EMT in a computationally affordable but still functionally rich manner. In their model Steinway et al. already involved β-catenin as a translocating protein, but they did not examine the role translocating proteins play in signalling processes and their underlying compartment-specific

functions systematically. Although Boolean models are suitable to assess these questions, no previous work addressed this topic from a systematic point of view.

The work of Steinway et al. served as baseline for our EMT related research, thus in this work I will usually refer to their simplified network model of 19 nodes and 70 edges as the “original EMT model” whereas our own EMT model will be named as a “compartmentalised EMT model”.

Objectives

This work can be divided into two complementary parts. During the first part we created a database of human translocating proteins called the Translocatome and then in the second part utilizing data in the Translocatome database we implemented a compartmentalised *in silico* Boolean model.

During the creation of the Translocatome database we wanted to create a database of human translocating proteins that extensively collects data on the translocation probability of human proteins. Thus we aimed to create a

framework which enables information collection about translocating proteins, and to incorporate those data into a database that can be accessible via the internet and also supports the addition of future information. Experimental procedures can depict protein translocations in a very complex way but our aim was to also have an extensive coverage of human proteins. So we wanted to utilize a machine learning based prediction tool (XGBoost) which is able to classify proteins based on training sets. Overall we aimed to create a database of human translocating proteins which builds on a manually curated set of known translocating (and non-translocating) proteins, and based on the predictive power of those datasets we wanted to predict the translocation probability of an extensive part of the human proteome.

During the second part we focused on utilising the data available in the Translocatome database, to create an *in silico* compartmentalised Boolean model where we can study the effect of subcellular dynamics and specifically the role of translocating proteins in cellular signalling. The observation of EMT is a suitable option as the EMT

was previously analysed via systems biology approaches, so there were some studies as comparisons to our compartmentalised model. Moreover, EMT is generally a well-studied process so *in vitro* comparisons were also available and the translocation of certain factors of EMT was already proven. We wanted to understand the role translocating proteins play in a subcellular processes and how the compartmentalised functions govern certain cell processes from a dynamic perspective. In summary, our aim was to utilize the data of the Translocatome database and predict translocating proteins of the EMT, then after validating those translocations create a compartmentalised Boolean model, where protein functions can be represented compartment-specifically. Based on dynamic simulations we wanted to analyse the compartment-specific functions of translocating proteins.

Results

Description of the Translocatome database

The Translocatome database is based on a manually curated core dataset containing 213 manually curated

human translocating proteins, which were collected via literature research of papers containing experimental evidence of protein translocations. Altogether the Translocatome contains 13 066 human proteins. The application of the gradient boosting machine learning tool, XGBoost enabled the prediction of translocation probabilities. This resulted in 1133 high-confidence translocating proteins, but all the 13 066 proteins of the Translocatome were characterized with a translocation likelihood, named as Translocation Evidence Score (TES).

Content of the Translocatome database

Translocatome is a database that contains 13 066 human proteins, each protein is characterized with a translocation probability and the core of the database is a strongly manually curated dataset. The interactome data of these proteins are also imported from the ComPPI database adding 151 889 protein-protein interactions.

For each of the 213 manually curated proteins in the database the following data (if the data were available) was collected:

- name set, gene name and UniProt accession number and link,
- PubMed ID(s) and link(s) to peer-reviewed article(s) describing the experimental evidence of translocation,
- initial and target localizations of the translocating protein,
- interacting partners and biological functions (both in the initial and target compartments),
- translocation mechanism,
- the used detection method,
- protein structural information on translocation mechanism,
- disease group, exact disease involved and pathological role,
- signalling pathways affected.

Predicting protein translocations with the Translocatome database

We utilized the well-established machine learning algorithm, XGBoost, to assess to translocation probability of the proteins in our database. Based on

positive and negative learning sets (160 translocating and 139 non-translocating proteins respectively) the algorithm was trained to predict TES values of proteins. The model performs with the average AUC of 0.9207 and both the precision-recall and Matthews correlation coefficient curves validated the accuracy of this prediction model. The model predicted 1133 high-confidence translocating proteins and 3268 low-confidence translocating proteins. The remaining 8665 proteins were classified as non-translocating.

The p63 protein (Tumor protein 63) was classified as a high-confidence translocating protein but it wasn't classified as a translocating protein based on the available literature. Our targeted literature review uncovered that p63 is a transcription factor (physiologically residing in the nucleus of human cells) and it plays an important role in the regulation of embryogenesis. Beside this p63 also appears in the cytoplasm of adenocarcinoma or prostate carcinoma cells, and this cytoplasmic localization results in increased malignancy of these tumours. This means that

our prediction of p63 as a translocating protein is valid and further *in vitro* studies could uncover the details of the translocation.

Enrichment of translocating proteins between signalling proteins

We have demonstrated, that the percentage of translocating proteins between human signalling proteins is higher (31% are low-confidence and 15% are high-confidence translocating proteins) than what we would expect based on a random sample. This deviation is significant ($p < 0.0001$, Chi-square test) and also true for the proteins of EMT.

Compartmentalised Boolean model of the epithelial-mesenchymal transition

The original EMT model had 19 nodes and 70 edges. Based on the data of the Translocatome 14 nodes out of the 19 were predicted as high-confidence translocating proteins. Compartmentalised Boolean rules were created based on available literature data; 64 publications were reviewed for the 10 compartmentalised nodes and thus by node-duplications we extended the 19-node network into

a 30 node network (the NOTCH node was divided into 3 subcellular nodes, since it has validated localization and activity in the plasma membrane, in the cytosol and in the nucleus as well). Our compartmentalised EMT model has seven attractors. Its most frequently occupied attractors are the epithelial (E) and mesenchymal (M) attractors. The analysis of the control sets of our model revealed that directing the network into the epithelial state is more feasible (requires less complex interventions) if we rely on compartment-specific perturbations.

Following the logic of our work compartmentalised models of biological processes can be created. We have developed a website which aides this process and enables the automated recreation and modification of our model.

Compartment-specific functions of translocating proteins

Our simulations showed, that in the presence of active TGFB signalling GSK3B were able to repress EMT, and this inhibitory effect is stronger if we perturb the nuclear pool of GSK3B. Similarly, GLI also has a compartment

specific role, as only the activation of GLI transcription factors in the nucleus is able to induce EMT.

Analysing signalling pathway activities via network models

We verified the known signalling pathway crosstalks with our *in silico* model and showed that Wnt and Hedgehog pathways, TGF-β and Hedgehog signalling pathways and TGF-β and Notch signalling all synergistically work together during EMT.

Conclusions

These investigations significantly moved us closer to understand the role of translocating proteins in protein-protein interaction networks and to enable further system level studies to be carried out. Our new findings are the following:

1. We collected and characterized a gold standard set of 213 translocating proteins and 139 non-translocating proteins. Using the XGBoost machine learning algorithm a Translocation Evidence Score (TES) was assigned to 13.066 human proteins.

Based on the TES values we have predicted 1133 and 3268 high- and low-confidence translocating proteins, respectively.

2. We created the Translocatome database as a user-friendly tool showing these data. We demonstrated the use of the Translocatome database by the p63 protein, which wasn't marked as a translocating protein in available databases, but after targeted literature review we could validate that it has roles in different organelles and cytoplasmic p63 expression serves as a biomarker in prostate cancer.

3. Translocating proteins are enriched between signalling proteins and the same enrichment is observed in the case of the proteins of epithelial-mesenchymal transition (EMT).

4. We built a compartmentalised Boolean model of EMT and developed a user-friendly tool, for an automated workflow of the process.

5. Our *in silico* dynamic simulations on the compartmentalised model showed, that GSK3B compartment-specifically inhibits EMT and GLI

transcription factors compartment-specifically drive EMT.

Bibliography of the candidate's publications

Publications directly related to the thesis

Mendik P, Dobronyi L, Hari F, Kerepesi C, Maia-Moco L, Buszlai D, Csermely P, Veres DV. (2019) Translocatome: a novel resource for the analysis of protein translocation between cellular organelles. Nucleic Acids Research, 47: D495-D505.

Mendik P, Kerestely M, Kamp S, Deritei D, Kunsic N, Vassy Z, Csermely P, Veres DV. (2022) Translocating proteins compartment-specifically alter the fate of epithelial-mesenchymal transition in a compartmentalized Boolean network model. npj Systems Biology and Applications, 8: 19.

Publications indirectly related to the thesis

Csermely P, Kunsic N, Mendik P, Kerestely M, Farago T, Veres DV, Tompa P. (2020) Learning of signaling Networks: molecular mechanisms. Trends Biochem Sci, 45: 284-294.