Analysis of system biology of cell compartments as a new test method of cancer development

PhD Thesis

Dániel Veres MD.

Semmelweis University

Doctoral School of Molecular Medicine

Supervisor: Prof. Péter Csermely, member of the Hungarian Academy of Sciences

Official evaluating committee: Dr. Bödör Csaba, PhD.

Horváth Zsolt, MD. PhD.

Chairman of the Comprehensive Exams Committee:

Prof. Edit Buzás

Members of the Comprehensive Exams Committee:

Dr. Miklós Cserző, PhD.

Dr. Attila Reményi, PhD.

Budapest

2017

Preamble

Neoplastic diseases belong to the leading death causes around the world, taking over the primary place of mortality from the former leading cause of death in some countries in the form of cardiovascular diseases. System-wide approaches (like e.g. network analysis) provide the opportunity of much more extended analyses using much more data than earlier, e.g. the pharmacogenomic analysis of *in vitro* cell line and *in vivo* human tumor molecular samples, which may help the more effective fight against tumors.

High throughput experimental methods provide a large amount of data related to protein interactions. Beyond increasing protein-protein interaction data it is essential to improve their reliability. The more reliable the data, the more precise biological hypotheses can be formed with analyzing them. There are lots of different approaches how to improve data quality. Data integration efforts stand out, which increase the probability of the existence of interactions derived from different tests and sources higher, than the biological probability of interactions based on single tests – that can be found in individual data bases.

The spatial position of proteins assures that they can fulfill different tasks in time and space, which is an important organizing mechanism of cellular biochemical processes. An expressive example of the compartment-level regulation of signal transmission paths is the activation of transcription factors mediated by their translocation to the cell nucleus. We define the concept of protein translocation as a systems biology phenomenon, where the proteins are relocated between two cell compartments, and where their translocation is regulated by post-translational modification. During translocation both the interacting partners of the protein and its function are changing.

Multifunctional proteins often change their locations via translocation to become able to perform their diverse functions. Mislocalization of proteins within the cells, e.g. by a pathological balance of translocation processes, may shift the healthy cell behavior into pathological direction, potentially contributing to strengthening or maintaining other pathological processes. One example of this is the ERK2 mitogen-activated protein kinase that is a typical multifunctional protein. ERK2 has different functions and biochemical activity in the cytoplasm and in the cell nucleus which are regulated by ERK's phosphorylation. This regulation is closely correlated with the extent of the malignancy of the tumor cells. So ERK2 is a good example of the important role that subcellular localization plays in protein function and signal transmission, and also of how important subcellular localization can be in the judgment of tumor progress' prognosis.

Balance of various regulations, related to proliferation and invasive behavior is of key importance in evaluation of tumors' malignancy. One of the causes of limited sensitivity towards conventional anticancer therapies or a developing resistance can be the presence of aggressively malignant cells proliferating slowly in tumors. For the treatment of such tumors a potential solution can be the two-step therapy targeting both conventional oncogenic pathways and metastasis formation.

Objectives

The objectives of my PhD thesis were the following:

1. **Creating a compartment-specific protein-protein interaction database:** The ComPPI database described in the thesis is a database and web portal that contains high quality compartment-specific protein-protein interaction data, integrated from numerous sources, of 791,059 interactions of 4 species' (yeast, worm, fruit fly and human) 125,757 proteins. The database predicts the subcellular localizations and the probability of interactions happening in the same compartment with separate scores.

2. **Creating a database based on manual curation and computerized data acquisition for collecting and predicting translocating proteins:** Translocatome is a database and an interactive web surface under development that is based on ComPPI screening and manual data collection, and contains detailed data, not summarized anywhere else yet, on subcellular localization and function of translocating proteins. These data are suitable for training a machine learning computer algorithm with the aim of predicting new translocating proteins.

3. **Better understanding of tumor initiation and progression with the help of network-biological observations:** The change of subcellular localization of proteins plays a key role in the development and the progression of tumors. The proteome- and interactome-level data of the databases developed during my doctoral work opened an opportunity to have a system-level examination of proteins related to tumor signal transmission, which helped to work out hypotheses on the development of tumor processes and interpret both former and newly assembled epidemiological data.

Methods

Selection of model organisms - ComPPI database contains data of the following four species: *Saccharomyces cerevisiae* – yeast-fungus, *Caenorhabditis elegans* – worm, *Drosophila melanogaster* – fruit fly, *Homo sapiens* – human.

Source of the protein localization and interaction data - ComPPI database summarizes 9 protein-protein interaction databases and 8 subcellular localization databases. Integration of the data sources help increase the amount of data and improve their quality because it decreases the possibility of data loss derived from the low overlapping of the databases. Subcellular localization data can come from experimental, predicted or unknown sources. ComPPI database contains physical interactions coming from experimental sources exclusively, which ensures the possibility of the functional analysis of relations. Experimental sources of interactions can be either low or high throughput experimental techniques.

The protein-protein interaction and subcellular localization data of Translocatome database are all derived from the ComPPI database. During the still ongoing manual curation of the translocating proteins PubMed and Google Scholar servers are used to search for articles describing proteins. For the identification of certain proteins and finding further functions mainly UniProt, NCBI Gene and GeneCards pages are used.

Methods used for functional analysis of proteins – For the annotation of proteins Gene Ontology database was used that was browsed through AmiGO web server. For the functional analysis of proteins BiNGO (3.0.2 version), the snap-in application of Cytoscape network visualizing and analyzing program capable of visualizing interactomes, was chosen. It is capable of the direct and interactive examination of the visualized network's elements. During building the Translocatome database the localization data are recorded according to the nomenclature of ComPPI, while the signal transmission paths and the disease names are systematized according to KEGG data source.

Tools used for network visualization and analysis of proteins – For visualization of networks two tools were used, one of them, Gephi 0.8.2 beta version was applied for quick visualization of larger networks. Cytoscape application was used for the visualization of subcellular localization-specific interaction network of certain proteins (also using the EntOpt network visualizing plugin developed by our research team), and for data analysis. NetworkAnalyzer tool that is built in the application and ModuLand network modulating algorithm developed by our team were used for calculating network indices.

Methods used for statistical analysis and data visualization – During the calculation and optimization of ComPPI localization and interaction confidence score and the general statistic data analysis and also for generating graphs R program package was used (3.1.0 version). For the sake of easier handling the R code was written and run with the help of the user-friendly RStudio application. Basic chart operations and statistical analyses were accomplished by the 2013 version of Microsoft Excel.

Methods for creating database and web surface - ComPPI source data are linked to the database with the help of interfaces. The database itself is MySQL 5 Community Edition based with its own nomenclature. The web surface was written in PHP5 language and based on nginx HTTP server. Translocatome database is MongoDB based, while the web surface helping the manual data collection is Ruby on Rails based. For machine learning scikit-learn package is used which is based on Python programming language.

Results

Schematic introduction of ComPPI database building process - A quite important application of ComPPI database is the possibility to screen the biologically not probable protein-protein interactions, where the two interacting partners have no common subcellular localization. The database also provides the possibility to predict new, localization-based functions of cellular proteins. In order to have a more extended amount of reliable data, the final ComPPI database was formed by the integration of a number of input databases. This process was manually controlled or assisted at quite a lot of occasions.

The demand for data integration arose during the first explorative data integration steps where a small overlap was found between the different databases. As a result of the extended search and control of the available data sources 9 protein-protein interaction databases and 8 protein subcellular localization databases were integrated to the ComPPI dataset.

Subcellular localization databases contain data having different levels of localization precision, e. g. predictive algorithms assign proteins only to large organelles, while experimental data provide information on the precise location of the proteins within the cells and their subcellular compartments. Besides the breakdown of the localization data, the compartments' names were also different in different source datasets. This required, as a first step, to change the various names of compartments to a unified nomenclature. The second step was to create a hierarchic, redundancy-free localization tree with manual collection that ranks the localizations unequivocally according to the Gene Ontology nomenclature.

Different input data sources often use various protein nomenclatures that need to be unified. For this purpose, an algorithm was developed by our team that, with the help of manually controlled translation boards, integrates the various names to the ultimate, more reliable UniProt names. After the successful integration of the data, the confidence score was calculated for both the localization and the interaction data.

The surface serving the database was tested in many ways with checking the search results, the downloaded data and the web functions.

Calculation of the confidence scores - ComPPI localization confidence score is a new index defined by our team with which the probability of certain proteins' location within a certain larger localization can be evaluated. The ComPPI localization confidence score can be used for building high reliability interactomes by combining it with the distribution of localization-specific interaction indices. The numerical value of the ComPPI localization confidence score depends on the localization data's experimental, predicted or unknown origin, as well as on the number of data sources.

The ComPPI interaction confidence score refers to the given interaction's subcellular localization-specific probability, whose basis is the consensus of the compartment-specific localization confidence scores of the interacting partners. The interacting partners' compartment-specific localization confidence scores, referring to six large localization categories, were calculated separately as the product of the respective, compartment-specific confidence scores. The ultimate interaction confidence score was calculated from the compartment-specific localization confidence scores of the six large localization categories.

During the calculation of the ComPPI localization confidence score, the source data's origin, i.e. the experimental, the predicted or the unknown evidence parameters are to be weighted in order to set up a reliable, unified scoring system for the diverse database. For weighing the evidence parameters data-driven optimization was applied.

Statistics of ComPPI database – The comprehensive and integrated database of the 1.1 version of ComPPI contains 127,757 proteins, their 791,059 interactions and 195,815 location data of the 6 large localizations. The proteome-level database includes localization data referring to the 5 large inter-cellular organelles (membrane, mitochondrion, cell nucleus, plasma cell, secretory path) and the extra-cellular compartment as categories. More than 60 percent of the localization records is of high resolution, covering over 1600 Gene Ontology cell components.

Introduction of the user interface of ComPPI – With the help of the web-based user interface of ComPPI (http://comppi.linkgroup.hu/) researchers not having bioinformatics skills can also browse the localization and interaction data of proteins of the dataset in a user-friendly way. Search function is helped by an automatic completion of the written protein's name decreasing the search errors caused by mistyping. Beyond simple search options there is an opportunity for extended search with which species, localization and localization confidence score specific searches can also be run. Filters can be applied for the searched protein and its interacting partners, too. There is an opportunity to download certain data sets using the ComPPI download page.

Usage of ComPPI data on certain cells' level – As the first step of our inquiry on the importance of those proteins, whose interaction is not biologically probable due to their different subcellular localization, a system-level search was implemented. We searched for such proteins in the human data set, whose interactome showed the largest change after filtering out the non-probable interactions. For this the number of interacting partners of the proteins was calculated and then the degree-distributions of the whole and the highly-reliable, filtered interactomes were compared. After this the first 20 proteins, showing the largest difference in their degree and betweenness centrality before and after filtering for high-probability interactions, were checked manually in the reliable UniProt SwissProt nomenclature hit list. Among the 20 proteins enoil-CoA hydratase (or in other, more conventional name: crotonase) showed the largest absolute change in degree, so this protein was selected for further analysis.

Crotonase is the enzyme catalyzing the second step of beta-oxidation of fatty acids that is one of the main members of the crotonase protein super-family. Beta-oxidation of fatty acids takes place mainly in the mitochondrion, which agrees well with the experimental data showing the mitochondrial localization of crotonase. According to cumulated data crotonase had 71 interacting partners. However, only 8 of them had mitochondrial localization, while only 5 had an interaction confidence score above 0.8 in this localization. Out of the 71 interacting partners 52 had localization data connecting them to the cytoplasm. Out of the 8 mitochondrion partners 7 had proven cytoplasmic localization, whose localizations' confidence score was above 0.8. Based on the above it arose that crotonase may have cytoplasmic localization as well, which would explain its numerous cytoplasmic interactions.

In order to prove the above assumption we searched the literature. The presence of crotonase in the cytoplasm of liver cancer cells was observed, where it played a role in lymphoid metastasis development. Following this path, the enrichment of the interaction partners' biological processes, recorded in the Gene Ontology database, was examined with the help of BiNGO. During the analysis of mitochondrial interactors, those processes enriched significantly, which were related to catabolism and negative regulation of apoptosis. In compliance with the information above it was shown that crotonase was upregulated in a number of tumor types and in case it was knocked out, the capacity of cells to survive in liver tumors decreased, and their apoptotic answer to cisplatin treatment increased. A similar role of crotonase was shown in mammary tumor cell lines where its downregulation facilitated the PP2-induced apoptosis.

Based on these findings it can be hypothesized that the high number of crotonase interactions, that appear like improbable (based on different subcellular localization of the interaction partners), may be the outcome of the transient and dynamic cytoplasmic subcellular localization of crotonase. In this localization crotonase takes part in the inhibition of apoptosis. Subcellular localization-based functions thus may unravel new anticancer intervention possibilities like the inhibition of crotonase's cytoplasmic anti-apoptotic function in liver or mammary tumors.

The system-level translocation database based on ComPPI data – Translocating proteins are important elements of compartment-level interactome with a necessary regulatory function and often with a pathological role. Our new database that contains translocating proteins and their interactions is being developed with ComPPI data. This database is named Translocatome.

As there was no comprehensive data set deciphering which proteins are translocating and which ones are not, our primary target was to collect translocating and non-translocating proteins and predicting new ones. According to system-biological definition, the probability of translocation is determined by the role of the translocating protein in the network and its biological functions. For this it is necessary to know proteins' interaction network and also to assign the biological functions of the certain proteins.

In order to define the probability of proteins' translocation systematically we need a reliable, manually curated data set that contains surely translocating (positive data set) and surely not translocating (negative data set) proteins to help the identification of those proteins, which have similar characteristics. For this literature-based protein characteristics are recorded in detail, so they can be used later for cross-checking the results and for the thorough analysis of the role of certain translocating proteins.

Using the negative and the positive data set and machine learning methods we have the opportunity to assign a translocating probability to proteins contained in the ComPPI dataset using both the interactome network parameters and the proteins' biological functions. Functional parameters are assigned to the proteins using the data of Gene Ontology. During machine learning several algorithms were tried. According to our results so far, the neural network method classifies translocating proteins the best (having over 80 percent accuracy). The precision of the algorithm can be improved by increasing the learning and validating data sets, and by the analysis of the biological sense of characteristic parameters chosen by the neural network method.

Data of the Translocatome database contain manually collected translocating and non-translocating proteins and their characteristics, a protein list coming from the ComPPI database and their predicted translocation confidence score, interactions among proteins, and the Gene Ontology-based assigned protein functions. Translocatome database can be browsed on or downloaded from the web interface http://translocatome.linkgroup.hu/, which is under construction. Besides providing user-friendly browsing and downloading options, the web interface of Translocatome database, unlike ComPPI, will give the opportunity for public data development. For this purpose a web tool has been developed by our team which helps manual data collection, control and with handling user privileges it enables data editing for several users at the same time.

Role of spatial location of proteins in defining tumor malignancy – According to the hypothesis integrating our own results and the available literature tumors' malignant transformation can be interpreted as a two-stage process, where the molecular network first transforms from the initial status to a more flexible and more plastic one in the early-stage tumors and then it gets back to a new, more rigid structure which stabilizes the late-stage tumor phenotype. According to the hypothesis during the adaptation of tumors the phenotype having a more plastic molecular network can be matched with fast proliferative and less invasive characteristics, while the more rigid network is more specific to the less proliferative resting-cells, which can specifically be more invasive.

Our hypothesis of two-stage development of tumors is also confirmed by several molecular observations that are connected to proteins' subcellular localization. The example proteins, which were introduced in my thesis, came from the development and manual data collection of ComPPI and Translocatome databases. The literature assigns subcellular localization-specific function to them in different stages of cancer progression such as high proliferation and metastasis. We also predicted a localization-specific functional difference to these proteins during our work that can be important in relation with this proteins' role in the development of tumor malignancy. As specific examples, MPS1 and ERK2 kinases help malignant proliferation with different functions depending on their subcellular localization. FAK1 kinase, hTERT telomerase, NANOG transcription factor, P53 tumor suppressor and ZEB1 transcription factor may help tumor cells' proliferation and/or metastasis depending on the context of their subcellular localization. These molecular phenomena were used by our team to help the molecular-level interpretation of American, Norwegian and Swedish colleagues' newly assembled colon tumor epidemiological data.

Conclusions

During my doctoral work I examined the role of proteins' spatial location in the protein-protein interaction network, which is an important prerequisite to maintain the proper signal transmission in healthy cells. In tumor cells the spatial localization of several proteins changes, which causes key changes of signaling promoting tumor progression as described by several examples of my thesis. Cell compartment-level analysis of protein-protein interactions provides the possibility for spatial analysis of the important signal transmission processes, e.g. for the system-level prediction and analysis of the functional consequences of protein translocation.

Main new results introduced in the paper:

1. We have created the first compartment protein-protein interaction database, ComPPI (http://comppi.linkgroup.hu/), which is one of the largest user-friendly interaction and localization data sources.

2. We have created Translocatome database (http://translocatome.linkgroup.hu/), which is the first systematically collected and annotated data set of the proteins that play roles in signal transmission regulation via translocation in human cells. We have collected examples for the roles of compartmentalization and translocation in malignant mutations.

3. Based on our network studies and experiences we have created the two-stage hypothesis of malignant transformation. First from the healthy cell a plastic pre-malignant network is formed that is able to react to the varying environmental effects. Then the molecular network becomes more rigid and stabilizes the new, often metastatic phenotype. In a collaboration paper using the proteins which change their functions by tumor-induced subcellular localization differences helping the interpretation of the newly assembled colon tumor epidemiological data of American, Norwegian and Swedish researchers we have summarized that the above hypothesis may have clinical importance at tumor prevention screenings, at planning anti-cancer therapy and also during the follow-ups.

List of my own publications

Papers in connection with the dissertation

1. Adami HO, Csermely P, **Veres DV**, et al. Are rapidly growing cancers more lethal? *Eur. J. Cancer.* 2017;72:210-214. (IF: 6,163, Google Scholar citations: 0)
2. Csermely P, Hódsági J, Korcsmáros T, (**Veres DV**) et al. Cancer stem cells display extremely large evolvability: Alternating plastic and rigid networks as a potential mechanism. Network models, novel therapeutic target strategies, and the contributions of hypoxia, inflammation and cellular senescence. *Semin. Cancer Biol.* 2015;30:42-51. (IF: 9,955, Google Scholar citations: 39, Web of Science citations: 20)
3. Gyurkó DM, **Veres DV**, Módos D, Lenti K, Korcsmáros T, Csermely P. Adaptation and learning of molecular networks as a description of cancer development at the systems-level: Potential use in anti-cancer therapies. *Semin. Cancer Biol.* 2013;23(4):262-269. (IF: 9,143, Google Scholar citations: 16, Web of Science citations: 12)
4. **Veres DV**, Gyurkó DM, Thaler B, et al. ComPPI: A cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic Acids Res.* 2015;43(Database issue):D485-D493. (IF: 9,202, Google Scholar citations: 21, Web of Science citations: 14)

Papers not in connection with the dissertation

1. Csermely P, Sandhu KS, Hazai E, (**Veres DV**) et al. Disordered Proteins and Network Disorder in Network Descriptions of Protein Structure, Dynamics and Function: Hypotheses and a Comprehensive Review. *Curr. Protein Pept. Sci.* 2012;13(1):19-33. (IF: 2,326, Google Scholar citations: 55, Web of Science citations: 36)
2. Simkó GI, Gyurkó D, **Veres DV**, Nánási T, Csermely P. Network strategies to understand the aging process and help age-related drug design. *Genome Med.* 2009;1(9):90. (IF: 0,0, Google Scholar citations: 32, Web of Science citations: 22)
3. Pákó J, **Veres D**, Tisza J, Horváth I. COPD in terms of multimorbidity. *Orvosi Továbbképző Szemle.* XXII. évf. 11. szám 2015.