Assembling and analyzing complex regulatory networks

PhD Theses

Dénes Türei

Semmelweis University
Doctoral School of Molecular Medicine

Supervisor:	Péter Csermely, professor, member of the Hungarian Academy of Sciences
Opponents:	Eszter Ari, assistant lecturer, PhD Miklós Cserző, senior research fellow, PhD
Chair of the examining board:	Sára Tóth, associate professor, PhD
Members of the examining board:	Krisztina Káldi, associate professor, PhD Csaba Hetényi, senior research fellow, PhD

Budapest
2014

1 Introduction

Amount of available molecular biological data and bioinformatic tools expanding more and more rapidly. The unmanageably huge amount of data implies the necessity of bioinformatics, and the demand for mathematical predictions, to support the planning of further research. Interaction databases consist a special class in the multitude of molecular biological data sources. These are collections of interactions most commonly between proteins, and sometimes between other kinds of molecules (RNAs, lipids, small molecules). We can regarde interaction databases as networks, where each molecule represented by a node, and each interaction by an edge in a graph. This abstraction gives us the advantage to analyse biological systems using the mathematical methods of graph theory.

In my PhD thesis I present three molecular interaction databases what I have built in the recent years. The first database is called SignaLink 2, and it aims to collect the connections and regulation of seven signaling pathways. Secondly I built a database about the regulation of NRF2, a transcription factor having key role in oxidative stress response and in numerous diseases. The third database offers a map of the regulation of autophagy, a cellular process highlighted in cancer, aging and neurodegenerative diseases.

The three new databases share a set of common design principles and methodology. Most importantly, the integration of the different levels of molecular regulation into one uniform data model: data on transcriptional, post-transcriptional and post-translational regulation are handled the same way, constituting a global cellular regulatory network. Another common feature is that the databases are based on original manual curation of the literature, which provides a good quality, trustable knowledge about the protein-protein interactions. The third common element is the signal transduction network, which connects the communication and information processing subsystems of the cell with the regulation of the examined processes.

All the three bioinformatic resources presented in my thesis, are gap-filling in their fields, as there is no other resource available where all these data would be accessible in a unified system – which is is a prerequisite of using them in analysis. However it would be misconception to consider them only as already existing databases now put together.

The data integration is only one step towards our aims. All the three databese are built around our own literature curation datasets. Members of our group collected information about protein-protein interactions from original publications. This way from thousands of articles – representing similar number of experiments – we compiled a protein-protein interaction network. Moreover, after selection of high-quality resources, we converted the different formats into our standard. We also added predictions and confidence scores to our databases. Some of them has been computed by us, and some has been taken from other databases.

2 Objectives

Considering the recent needs in the field of analysing molecular regulatory networks, and aiming a convenient utilization of the data provided by modern experimental technics, I set the aim to extend the SignaLink 1 signaling database in multiple directions. This extension includes the following elements:

- enrichment of the SignaLink 1 literature curated signaling network with post-tanslational regulation data from other databases
- prediction of further post-translational regulators, inferred from domain-motif interaction data
- assignement of directions to undirected interactions of high-throughput experiments, where it can be predicted from domain-domain interaction data
- collecting transcriptional and post-transcriptional regulators of the proteins involved in the previously described network, including the transcriptional regulation of miRNAs
- to design an efficient and extensible database structure to store the described data
- to provide modern and interactive web interfaces for browsing the contents of the databases
- to make the data easily accessible for bioinformatic analysis in multiple standard formats by a customizable download interface.

The aim of the second part of my PhD was to build the regulatory network of NRF2, a highly important transcription factor. To realize this, I set the following aims:

- to create a database based on literature curated data, including post-translational, transcriptional and post-transcriptional regulation
- to connect this NRF2 focused regulatory network with the signaling pathways from SignaLink 2
- to make the data browsable and downloadable by a webpage and a download module.

The third part of my PhD, using similar methodology, I had the aim to map the regulation of autophagy, a cellular process in the focus of numerous current research, raising several unanswered questions, and playing key role in the pathomechanisms of many diseases. To achieve this, I set the following aims:

- to combine the manually collected network of autophagy proteins with post-translational regulators from other databases
- to include transcriptional and post-transcriptional regulators, and connect them with signaling pathways from SignaLink 2
- to provide a webpage and download interface for this database.

3 Methods

All the three databases presented in my thesis are built around curation of literature. In the case of ARN, it has been done by László Földvári-Nagy, member of our group, with the aim to create ARN. The other two databases, the SignaLink 2 and the NRF2ome are based on previous works of members of my research group and my consulent, Prof. Péter Csermely.

The starting point of SignaLink 2 development is the SignaLink 1 database. SignaLink 1 consists literature curation of eight signaling pathways in three species, including pathway members from receptors and ligands to the first transcription factors.

The development of NRF2ome started from the NRF2 regulatory network. This network includes the direct regulators of NRF2 and KEAP1, having almost a star topology, and contains 289 protein-protein interactions, 7,469 genes transcriptionally regulated by NRF2, 63 miRNAs suppressing the translation of NRF2, and 35 transcription factors regulating those miRNAs. The data is available as six separate tables in the supplement material of Papp et al. The PubMed IDs of the referred articles are stored in the database, and the references describing a connection between proteins are assigned to the interactions. Members of our group performed an extensive search in the PubMed database to collect studies describing interactions between proteins. In the case of SignaLink 1, all the eight pathways in the three species (*C. elegans, D. melanogaster and human*) has been collected the same way, but independently. In the course of constructing SignaLink 2 we updated this collection with new publications from the recent two years, and we extended with interactions between pathway members and endocitotic and scaffold proteins. In addition, we merged the MAPK and IGF pathways into one pathway, called RTK.

The core of NRF2ome consists one set of manually curated data, which covers the direct interactors of NRF2, some important interactions between these proteins, and the most important partners of KEAP1, which is the main regulator of NRF2. All together, this collection draws up a network of 108 proteins and 146 interactions.

Similarly, the ARN database is based on literature curation, describing the interactions between the 38 autophagy executor proteins (autophagic machinery), and some of their regulators.

Reference databases I mapped all the various protein identifiers to UniProtKB identifiers. The SignaLink 2 contains the the Ensembl identifiers of human proteins, the WormBase identifiers of *C. elegans* proteins, and the FlyBase identifiers of *D. melanogaster* proteins as well. The miRNA names have been mapped to miRBase pre-miRNA accession numbers.

Protein-protein interaction databases Considering the generic protein interaction databases, I included data from BioGRID, HPRD, InnateDB and IntAct.

Predictions Novel interactions between proteins have been predicted, based on domain-motif interactions, using the ELM Structure Filter service.

Predicted directions The waste majority interactions from generic protein-protein interaction databases has unknown direction. We tried to improve this by predicting directions using protein domain constitution data from Pfam, and domain-domain interaction data from DOMINE databases. We applied a machine learning algorithm, using Reactome data as gold standard positive, and Negatome as negative standard. Finally, we performed a ROC analysis to reduce the number of false positives.

Transcriptional regulation From ABS, DroiDB, edgeDB, ENCODE, RedFly, ORegAnno, PAZAR and wTF databases experimental TF–promoter binding data has been taken into our databases. Furthermore, we predicted novel TF binding sites applying JASPAR algorithm on the 2000 nucleotide long region preceding the transcription starting points.

Post-transcriptional regulation We fetched predicted miRNA–RNA interactions from microT-v4, miRanda, miRDB, PicTar and TargetScan algorithms, and experimentally verified interactions from miR2Disease, miRDeathDB and TarBase databases. Data on trancriptional regulation of miRNAs has been extracted from ENCODE, PuTmiR and TransmiR databases.

Properties of interactions In the database structure what I designed, interactions might be direct (e.g. phosphorylation), or indirect (e.g. transcriptional regulation); their effect might be stimulatory, inhibitory or unknown; and they can be directed or undirected. I assigned various confidence scores to interactions. In case of protein-protein interactions I calculated a semantic similarity score based on Biological Process attributes of proteins from Gene Ontology. For some of the human interactions I used the PRINCESS combined confidence score, which takes into account domain constitution, subcellular localization, expression pattern, genomic context and network topological data. In case of JASPAR prediction and predicted miRNA–mRNA interactions, the quantitative result of the prediction algorithm serves as a confidence score.

Informatic background I constructed the databases using MySQL database management system, running on Linux operating system. The webpages are built upon PHP, jQuery, jQuery UI and Cytoscape Web, while the export module has been written in Python.

4 Results

SignaLink 2 SignaLink 2 is a novel bioinformatic resource about the regulation of seven signaling pathways: Hedgehog, WNT/Wingless, TGF-β, RTK, JAK/STAT, Notch and NHR. The database contains the interactions between pathway components, and their post-translational, transcriptional and post-transcriptional regulators. SignaLink 2 offers a comparative map of this regulatory network from three species: *C. elegans,* and *D. melanogaster,* which are important model organisms, and human.

In SignaLink 2 the interactions between proteins and miRNAs are classified into layers, categories built upon each other in an onion-like manner. These layers represent the different ways of regulation observable in molecular systems. Furthermore, the layers differ from each other also in their data sources. This clear classification offers a convenient way for researchers using SignaLink 2 to make decision whether to include or not in their analyses the different levels of regulation, and data sources offering different amount of data with different levels of confidence.

NRF2ome In the construction of NRF2ome I followed the multi-layered design of SignaLink 2, with some modifications. The core of the database is similarly a manually collected dataset, containing the direct interactiors of NRF2 and KEAP1. Further regulators of the components of this network have been included from other databases, domain-domain and domain-motif based prediction methods. The following two layers contain the transcriptional and post-transcriptional regulators, from the same data sources as SignaLink 2. In NRF2ome, the last layer connects the signaling pathways from SignaLink 2 with the regulatory

Table 1: **Comparison of SignaLink 2, NRF2ome and ARN**

	SignaLink 2	NRF2ome	ARN
Interactions	363.998	36.139	263.411
Proteins	33.105	7.891	4.034
miRNAs	872	541	1.380
Data sources	59	47	59
References	8.446	2.846	2.023

network of NRF2, both on post-translational, transcriptional and post-transcriptional level.

The Autophagy Regulatory Network (ARN) The architecture of ARN is mostly similar to NRF2ome, but here the regulatory network is built around not only one, but 38 proteins – those are involved directly in the execution of autophagy.

4.1 Az adatbázisok mennyiségi tulajdonságai

The database webpages All the three databases have their own webpage. SignaLink 2 is accessible under http://signalink.org/, NRF2ome at http://nrf2.elte.hu/, and ARN at http://arn.elte.hu/. These webpages give an opportunity to browse the data interactively, and to use the download module. With the intelligent name search function finds makes possible to find proteins and miRNAs searching by various names and database identifiers. Protein and miRNA datasheets show a list of interactions grouped by layers. Sources, references and confidence scores of the interactions can be viewed a comparative way. An interactive visualization of the first neighbours of the selected component is shown on the datasheets. The download module is available throught the webpage, where users can easily select and combine the species, pathways, layers and the file format of download. Data can be exported in major bioinformatic standard formats, including csv, BioPAX (level 3), PSIMI-TAB, PSIMI-XML, SBML and Cytoscape.

5 Conclusion

Availability of data on post-translational, transcriptional and post-transcriptional regulation rapidly expanding, thanks to the development of molecular biological technologies and mathematical prediction methods. Several bioinformatic resources have been created to provide these data for the scientific community. The different, sometimes overlapping content of the databases raise difficulties in satisfying basic demands of research, e.g. collecting all known transcriptional regulators of one protein. Because the different quality of data obtained from various experimental technics and prediction methods, filtering data is often needed to fit the demands of analysis methods. Manual curation of literature provides the most trustable data, at the expense of tedious work, moreover well defined curation protocol is needed to maintain comparability.

With SignaLink 2 we provide a gap-filling resource to resolve these issues, integrating all levels of molecular regulation into a uniform interaction network, and at the same time, offering the opportunity to customize the dataset according to the current needs. The other two novel resources presented in my PhD thesis, NRF2ome and ARN draw maps on the regulation of two highly significant cellular process. NRF2 is the master transcription factor of antioxidant response, affecting the transcription of more than 7,500 genes. It has important role in the defence against xenobiotics, in cardiovascular diseases, inflammation and cancer. Autophagy also has significant functions in the patomechanism of numerous diseases, such as cellular damage caused by ischaemia-reperfusion, neurodegenerative diseases and tumors. The ambiguous role of NRF2 and autophagy in these diseases, described by the "double edged sword" metaphore, implies the challange in finding new therapies. Until now, there was no system-level resource available about the regulation of these processes, despite such resources are undispensable in the synthesis of our existing knowledge, discovering new relations, to explore the context dependent ways of regulation, and to model the effects of potential drug compounds.

SignaLink 2, NRF2ome and ARN offers a good basis to use various methods to analyse data. Depending on the modeling methods to use, the different questions can be answered, and different system sizes, on different levels of resolution. The differential equation systems or rule-based models are limited to precise, quantitative simulation of interaction networks below the size of 100 molecules. Modularization methods, network topology analysis and perturbational simulations are suitable to give an overall view on the properties of large networks. Combining the generic interaction networks presented in my thesis with gene expression or mutation data, tissue or cell line specific networks can be generated for further analysis.

The SignaLink 2, NRF2ome and ARN databases show a significant improvement in the field of bioinformatic data integration and construction of molecular regulatory networks. The informative and user-friendly webpages, and the download service offering data in modern standard formats, open the way towards the application of these recent bioinformatic resources in a broad range of researche.

6 Publications

Related publications

1. Papp D., Lenti K., Módos D., Fazekas D., Dúl Z., **Türei D.**, Földvári-Nagy L., Nussinov R., Csermely P. and Korcsmáros T. (2012). The NRF2-related interactome and regulome contain multifunctional proteins and fine-tuned autoregulatory loops. *FEBS Lett*, 586(13): 1795–1802. (IF: 3,54)

2. Fazekas D.*, Koltai M.*, **Türei D.***, Módos D, Pálfy M, Dúl Z, Zsákai L, Szalay-Bekő M, Lenti K, Farkas I, Vellai T, Csermely P and Korcsmáros T (2013). SignaLink 2 – a signaling pathway resource with multilayered regulatory networks. *BMC Syst Biol*, 7: 7. (IF: 3,15)

3. **Türei D.**, Papp D., Fazekas D., Földvári-Nagy L., Módos D., Lenti K., Csermely P. and Korcsmáros T. (2013). NRF2-ome: an integrated web resource to discover protein interaction and regulatory networks of NRF2. *Oxid Med Cell Longev*, 2013: 737591. (IF: 2,84)

4. **Türei D.**, Földvári-Nagy L., Módos D., Fazekas D., Csermely P., Vellai T. and Korcsmáros T. (előkészület-

ben). Autophagy regulatory network – an integrated resource to identify novel regulations and interactions that conrol autophagy.

* – shared first authorship

Other publications

1. Hufnagel L., Gaál M., Ladányi M., Cs S., Petrányi G., Aczél D., **Türei D.** and Zimmerman D. (2005). Klímaváltozás potenciális hatásai magyarország rovarfaunájára. *VII. magyar biometriai és biomatematikai konferencia: összefoglalók*. Ed. by Szenteleki K. and Szilágyi K.: 21.

2. Hufnagel L., Sipkay C., Drégelyi-Kiss Á., Farkas E., **Türei D.**, Gergócs V., Petrányi G., Baksa A., Gimesi L., Eppich B., Dede L. and Horváth L. (2008). *Klímaváltozás: környezet–kockázat–társadalom*. Ed. by Hufnagel L. Szaktudás Kiadó Ház, Budapest. Fejezet: Klímaváltozás, biodiverzitás és közösségökológiai folyamatok kölcsönhatásai.

3. Ferenczy A anf Eppich B., Varga R. D., Bíró I., Kovács A., Petrányi G., Hirka A., Szabóky C., Isépy I., Priszter S., **Türei D.**, Gimesi L., Á G., Homoródi R. and

Hufnagel L. (2009). Fenológiai jelenségek és meteorológiai indikátorok kapcsolatának összehasonlító elemzése rovar és növény adatsorok alapján. *LII. georgikon napok: gazdaságosság és/vagy biodiverzitás?* Ed. by Szenteleki K. and Szilágyi K. Keszthely: 35.

4. Vadadi-Fülöp C., **Türei D.**, Sipkay C., Verasztó C., Drégelyi-Kiss Á. and Hufnagel L. (2009). Comparative assessment of climate change scenarios based on aquatic food web modeling. *Environ Model Assess*, 14(5): 563–576. (IF: 0,97)

5. Ferenczy A., Eppich B., Varga R. D., Bíró I., Kovács A., Petrányi G., Hirka A., Szabóky C., Isépy I., Priszter S., **Türei D.**, Gimesi L., Á G., Homoródi R. and Hufnagel L. (2010). Comparative analysis of the relationship between phenological phenomena and meteorological indicators based on insect and plant monitoring. *Appl Ecol Env Res*, 8(4): 367–376. (IF: 0,38)

6. Verasztó C., Kiss K. T., Sipkay C., Gimesi L., Vadadi-Fülöp C., **Türei D.** and Hufnagel L. (2010). Long-term dynamic patterns and diversity of phytoplankton communities in a large eutrophic river (the case of

river danube, hungary). *Appl Ecol Env Res*, 8(4): 329–349. (IF: 0,38)

7. Hufnagel L., Kúti Z., Hlaszny E., Reiczigel Z., Molnár M., Homoródi R., Flórián N., Gergócs V., **Türei D.** and Ladányi M. (2012). A klímaváltozás közösségökológiai hatásainak elemzései. *Fenntartható fejlődés, élhető régió, élhető települési táj; tudományos közlemények III.* Ed. by Szenteleki K. and Szilágyi K. Budapesti Corvinus Egyetem, Budapest: 7–24.

8. Komoly C., **Türei D.**, Csathó A. I., Pifkó D., Juhász M., Somodi I. and Bartha S. (2012). Fűvetés hatása a parlagfű (*Ambrosia artemisiifolia* L.) tömegességére egy tiszaalpári fiatal parlagon. *Természetvédelmi Közlemények*, 18: 283–293.

9. **Türei D.** (2012). *A klímaváltozás hatása ökológiai folyamatokra és közösségekre*. Ed. by Hufnagel L. and Sipkay C. Budapesti Corvinus Egyetem, Budapest. Fejezet: Vízi és vizes élőhelyek specifikumai: 85–128.

7 Aknowledgements

Here I would like to express my gratitude to all my collegues and friends who have helped and supported me along my PhD research.

Foremost to my supervisor, Professor Péter Csermely, member of the Hungarian Academy of Sciences, to Professor Gábor Bánhegyi, the director of the Department of Medical Chemistry, Molecular Biology and Pathobiochemistry, Semmelweis University, and to Professor József Mandl, member of the Hungarian Academy of Sciences, head of the Doctoral School of Molecular Medicine.

Special thanks to all members of the LINK group at Semmelweis University, and the Network Biology group at Eötvös Loránd University, in particular to Tamás Korcsmáros, Dávid Fazekas, Dezső Módos, Diána Papp, László Földvári-Nagy, János Kubisch, Illés J. Farkas and Zoltán Dúl.

My reasearche was financially supported by the TÁMOP grants number 4.2.1./B-09/1/KMR-2010-0003 and 4.2.2/B-10/1-2010-0013 and OTKA grants number K83314, K75334 and NK78012 and the NKTH grant number 5LET-08-2-2009-0041.

The Semmelweis University Central Library provided me the access to scientific journals and literature.

I am beholden to the thousands of programmers contributing in the free software movement.

Finally I would like to thank my family, and the people of the Lujza utca commune: Hajnalka Bezgődi, Csilla Hódi, Sharon Kathrin, Emma Krasznahorkai, Sarolta Kremmer and Lujza for the endless energy I got from living together with them.