Ph.D. THESIS SUMMARY

System-level database and analyses of signaling pathways: identification of novel protein functions, signaling cross-talks and drug targets

Tamás Korcsmáros

Supervisors:
Prof. Péter Csermely
Semmelweis University
Department of Medical Chemistry

Dr. Tibor Vellai
Eötvös Loránd University
Department of Genetics

Dr. Balázs Papp
Biological Research Center of the Hungarian Academy of Sciences
Institute of Biochemistry

Doctoral School of Biology
University of Szeged

Szeged
2011

Introduction:

Intracellular signaling contributes extensively to the diversity of developmental programs and adaptation responses in metazoans. In humans, defects in intracellular signaling can cause various diseases, e.g., cancer, neurodegeneration, or diabetes. Thus, understanding the structure, function, and evolution of signal transduction is an important task for both basic research and medicine.

Signaling pathways transduce the information of various extracellular ligands to the nucleus via receptors, secondary messengers and transcription factors. In the nucleus specific changes in gene expression occur as a response to the external stimuli. Interestingly, the number of signaling pathways is low and each pathway is formed by only 10-20 proteins. Thus, these numbers are contradictory to the overall scale of cell types and responses that is governed by signaling pathways.

Currently, high-throughput (HTP) experiments are the major sources of known protein-protein interactions. However, so far in most HTP experiments extracellular, membrane-bound, and nuclear proteins have been underrepresented. These and other sampling biases strongly reduce their usability for identifying signaling interactions. Another limitation of HTP assays is that they produce undirected interactions even though in signaling directions are essential. Accordingly, several signaling pathway databases have been created recently by manually collecting the directed interactions from the literature.

Manually curated signaling pathway databases are often assembled without strictly defined and published standardized curation criteria. Therefore, even within the same database, e.g., in KEGG, the level of detail of curation and the rules for setting pathway boundaries can vary among pathways. In addition, in several signaling resources the definition of signaling pathways has no evolutionary or biochemical background. In other cases, e.g., in Reactome and NetPath, curation criteria are standardized, however, (i) pathways are usually handled as separate entities, (ii) cross-talks and multi-pathway proteins are underrepresented, and (iii) extracting signaling information from the databases are complicated and labor-intensive. Another limitation of several current signaling resources is that they neglect the importance of multi-pathway proteins, i.e., proteins functioning in more than one pathway. In summary, the manual curation process needs to be uniform across all pathways and species to aid

cross-talk analyses, tests of evolutionary hypotheses, dynamical modeling, setting up predictions, and drug target selection.

Intracellular signaling was originally regarded as an assembly of distinct and almost linear cascades. Over the past decade, however, it has been realized that signaling pathways are highly structured and intertwined with cross-talks (where cross-talk is defined here as a directed physical interaction between pathways). Consequently, intracellular signaling is now viewed as a set of intertwined pathways forming a single signaling network. This paradigm shift calls for novel experimental, curation, and network modeling techniques.

Aims of the work:

My aim was to compile a signaling pathway database that is:

- novel and can facilitate modern signaling research, including network approaches,
- based on well defined pathways and can be generally used.

Creation, visualization and validation of such a database requires:

- the selection of specific organisms and pathways, where adequate amount of information is accessible,
- a curation protocol, that is reproducible and objective,
- the development of novel visualization methods.

Beside the creation of this database, I intended to perform analyses that were not possible before. Thus, my aim was to illustrate and prove that the newly developed database:

- is a proper source for system-level analyses,
- can facilitate cross-species comparison of signaling pathways and prediction of novel protein functions,
- allows the analysis of cross-talks between signaling pathways and the identification of multi-pathway proteins,
- is an efficient source for applied research and development approaches (e.g., cancer research, drug development).

Methods:

SignaLink lists signaling proteins and directed signaling interactions between pairs of proteins in healthy cells of *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*. Each interaction is documented with the PubMed ID of the publication reporting the verifying experiment(s). SignaLink was compiled separately for the 3 organisms. SignaLink assigns proteins to signaling pathways using the full texts of 170 pathway-based reviews. Interactions were curated from a total of 941 articles (PubMed IDs are included in the database). We selected 8 major pathways for curation that have central roles in both development and normal cellular signaling. The curation protocol allowed a protein to be assigned to more than one pathway, thus we could list multi-pathway proteins.

Due to the absence of appropriate gold standards we compared the human signaling pathways of SignaLink with those from 3 widely used pathway databases: KEGG, Reactome, and NetPath. For reference we compared the human signaling pathways of these three databases to each other too.

We visualized the signaling pathways and cross-talk of SignaLink with traditional and novel visualization methods. This was necessary to illustrate and examine the signaling network. Altogether, we have applied 7 visualization methods based on the signaling, orthology and cross-talk properties of the network components.

In each of the three species examined, we listed those proteins that have no known signaling interactions but have at least one signaling pathway member ortholog in the other two species. Similarly to the concept of functional orthology, for each of these proteins we assumed that their pathway annotations (i.e., signaling role) can be transferred between species. In other words, we predicted that such a protein is a member of the signaling pathway(s) to which its ortholog(s) in the organism belong(s). These proteins were termed as signalog proteins (signalogs).

To investigate the dynamic activity of pathway interactions, we selected five healthy tissue types – colorectal, muscle, skin, liver, and cardiovascular tissues – and 2 liver carcinomas. Protein expression data in healthy tissue types were downloaded from the eGenetics database. Protein expression data in 2 screens of liver carcinomas were obtained from Oncomine 3.6. We considered a protein differentially expressed if the p

value of its expression in at least one of the 2 screens, as compared to healthy liver tissues and computed by a t-test of Oncomine, was below 0.05.

We collected information on the proteins that can be relevant in drug target discovery with DAVID. We downloaded disease-related annotations from OMIM, GAD, and Orthodisease, domain information from InterPRO, and molecular function and cellular component data from GO.

Results:

1. The SignaLink database

We curated the signaling pathways of the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, and *Homo sapiens*. From the wide variety of classification schemes for selecting signaling pathways we followed the biochemical approach of Pires-daSilva and Sommer. We selected 8 major pathways for curation – EGF/MAPK, Ins/IGF, TGF-β, WNT, Hh (Hedgehog), JAK/STAT, Notch, and NHR (Nuclear Hormone Receptors) – that have central roles both in development and in normal cellular signaling.

SignaLink is a manually compiled resource integrating experimentally confirmed genetic and physical interactions from healthy tissue types. Proteins and interactions are listed without tissue-specificity and can be visualized as networks of potential interactions. Tissue- and disease-specific information can be added easily as shown in the examples below. Five combined characteristics create the unique utility of SignaLink:

- Pathways are biochemically defined and encompass all major developmental signaling mechanisms;
- A protein can belong to more than one pathway (if it does, then it is called a multi-pathway protein);
- Proteins are tagged with (i) the pathway(s), (ii) pathway region(s) (core, peripheral), and (iii) the pathway sections (one or two of: ligand, receptor, mediator, co-factor, transcription factor, other) they belong to;
- The level of detail is the same for the entire database;

- Interactions are directed and manually labeled with PubMed IDs (experimental evidence).

The current version of SignaLink (published in 2010) contains 442, 211, and 525 proteins in *C. elegans*, *D. melanogaster*, and humans, respectively. Between these proteins it contains 237, 233, and 991 interactions, respectively. The SignaLink database can be freely accessed at http://SignaLink.org.

2. Comparing SignaLink database with other pathway databases

There are currently no gold standards compiled with similar goals and methods as SignaLink. It is therefore important to compare both the curation protocols and the actual data of several available databases before selecting one of them for a particular analysis. We compared three widely used pathway databases – KEGG, Reactome, and NetPath – and SignaLink. In each pairwise comparison we used the pathways available in both databases.

According to this comparison, SignaLink has the following advantages compared to the three analyzed databases:

- precisely defined and documented curation protocol;
- highest numbers of signaling proteins and interactions in the curated signaling pathways;
- highest numbers of cross-talks and multi-pathway proteins;
- largest protein overlap with the other databases;
- above the average number of publications used per pathway.

3. Visualization of the SignaLink database

Altogether, we have applied 7 visualization methods based on the signaling, orthology and cross-talk properties of the network components. Color was the main differences between the traditional, protein-based network visualization approaches (e.g., we colored the proteins and their interactions based on their pathway or signaling position property). We created 8-8 pathway images for the 3 species. All the network

images can be interactively explored at http://SignaLink.org. To visualize the cross-talk networks we applied 3 different approaches:

- A network visualization where each node represent a pathway (a set of proteins) and the edges between the nodes represent cross-talks (a set of protein-protein interactions). Coloring was done based on signaling or orthology properties.
- We created an interaction matrix, where specific cross-talks are present in each cell. In this case to color the cells, we used expression data of proteins.
- Finally, multi-pathway proteins were visualized: each node represents a pathway, and the edges between these nodes represent the overlaps between the pathways (not the interactions as in the first approach).

4. Results of the analyses performed with SignaLink

The newly developed SignaLink database has properties not present in other signaling pathway databases and its quantitative and qualitative features allow the system-level analysis of signaling networks.

4/a Identification and analysis of novel gene functions

We identified novel signaling pathway components based on the signaling pathway memberships of orthologs in another organism. We found 88, 92, and 73 proteins in *C. elegans*, *D. melanogaster* and *H. sapiens*, respectively, which had previously not been assigned to a signaling system, but have at least one ortholog in the other two species that is clearly associated with a signaling pathway. We hypothesized that these 253 proteins function in the same signaling pathways as their orthologs. Thus, we named the predicted signaling components signalog proteins, or briefly signalogs. The complete list of signalogs can be accessed at http://SignaLink.org. We examined the novelty of the signalogs, and verified their novelty.

4/b Comparison of signaling pathways within and between species

In all 3 organisms a few of the 8 pathways are central and abundant. Of all proteins 26% to 38% participate in the EGF/MAPK and WNT pathways, respectively. Other pathways with high protein numbers are NHR in the worm, Hh and Notch in the

fly, and TGF and JAK/STAT in humans. Altogether in each species 68% to 85% of all signaling proteins participate in these pathways and 56% to 70% of all cross-talks involve the EGF/MAPK, TGF, or WNT pathways. *C. elegans* has almost identical numbers of core and peripheral proteins in each pathway (except for Notch and NHR), while in the other two species the ratio of core to peripheral proteins is around 1.5.

Pathway size differences between the 3 species are often related to the different environments to which the cells of these organisms have adapted. For example, ligands from the environment can easily reach the nuclei of the worm's cells, thus, the worm's NHR pathway is exceptionally large (58% of all signaling proteins). On the other hand, due to the large variety of signals that human cells are exposed to the human JAK/STAT pathway is oversized com-pared to the other two species (21% of all signaling proteins in humans vs. 0% and 4% in *C. elegans* and *Drosophila*, respectively).

In all 3 species EGF/MAPK and IGF have high numbers of mediators. However, environmental differences may affect pathway section sizes too. In *C. elegans* transcription factors – dominated by the NHR pathway – are the largest pathway section (39%). In the other two species co-factors by far outnumber other pathway sections (32% to 42%) and in humans JAK/STAT ligands and receptors are abundant.

4/c Identification and analysis of multi-pathway proteins

In *C. elegans, D. melanogaster*, and humans, we found 6, 12, and 62 multi-pathway proteins, respectively. Within one human signaling pathway the ratio of proteins functioning in at least one other pathway varies from 5% (Notch) to 46% (IGF). Interestingly, a single protein can be even a central (i.e., core) component in more than one pathway.

We found that EGF/MAPK – the largest pathway – is the only one sharing proteins with all other pathways. On the other end of the spectrum are the Notch, JAK/STAT, and NHR pathways: their proteins are contained by 3 or 4 other pathways. These differences correlate well with the numbers of pathway functions. Note also that the set of 62 human multi-pathway proteins is enriched with disease-related proteins: 45% (28) of them are known to be disease-related, while in the 8 human signaling

pathways only 25.5% (165 of 646) and among all human proteins listed by Ensembl only 20% (3,929 of 19,534). For both comparisons $p < 0.001$.

4/d Cross-species comparison of cross-talks

In *C. elegans* only 6 of the 8 curated pathways are active, and the Notch pathway is isolated. In addition, the cross-talk network of the pathways – where nodes represent pathways and links represent cross-talks – is sparse. Between the 6 active pathways only 5 of the 30 (= 6×5) possible cross-talk types are present. In *Drosophila* all 8 curated pathways of SignaLink are active, but the NHR and JAK/STAT pathways are still isolated. Without these two pathways the cross-talk network is already significantly denser than in the worm: 16 of the total 30 possible cross-talk types are present. In humans – the most complex organism of the three – all 8 curated signaling pathways are active and almost all of the 56 possible cross-talk types are possible. The ubiquity of cross-talks (all 28 pathway pairs can cross-talk) expands both the repertoire of possible phenotypes and the system-level responses to environmental and pathological changes.

In *C. elegans* cross-talk is possible through receptors, mediators, and transcription factors. In the other two species all pathway sections can participate in cross-talk, except for the NHR and JAK/STAT pathways of *Drosophila*, where cross-talk occurs mostly at the transcriptional level and through mediators.

The presence of cross-talks in many pathways and pathway sections is a sign of the efficient utilization of resources: expanding the functions of an already existing pathway protein is more efficient than evolving a novel protein. In addition to the number of active pathways and cross-talks a further important indicator of signaling complexity is the number of cross-talks relative to all signaling interactions. In the worm 4.6% of all signaling interactions are cross-talks, in the fly 10.5%, and in humans 30.3%. Interestingly, the growth of the number of cross-talks from worm to fly and human is not simply due to the growth of the number of protein-coding genes (20 100, 13 800, 23 000, respectively) or the number of signaling-related PubMed articles (3 889, 11 367, 214 193 in worms, flies, and humans, respectively).

4/e Tissue- and disease-specific activity of cross-talks

Cross-talks, similar to other protein-protein interactions, are not active permanently in all tissue types. We considered an interaction to be possibly active in a given tissue type, if both of the mRNAs of its participating proteins are expressed in that tissue. The following pairs of large pathways the ratio of active signaling cross-talks is lower than the average: EGF/MAPK-IGF, IGF-JAK/STAT, EGF/MAPK-JAK/STAT, IGF-TGF, and IGF-WNT. In contrast for 3 larger (and several smaller) pairs of pathways cross-talk is more frequent in the investigated tissue types than one would assume from the sizes of the two cross-talking pathways: EGF/MAPK-NHR, NHR-TGF, and NHR-WNT. Cancers are often viewed as systems diseases. In cancer cells large-scale modifications of signaling pathways, especially changes of cross-talks, are prevalent. We considered a signaling interaction to be altered in liver carcinomas, if, compared to healthy liver tissues, at least one of the participating proteins was differentially expressed. In 3 of the 8 pathways (WNT, NHR, and JAK/STAT) only ~30% of all proteins were differentially expressed in the investigated liver carcinomas, while in the other 5 pathways this ratio was ~50%.

Finally, we concluded that the pathways EGF/MAPK, JAK/STAT, and Notch are clear examples for three distinct types of signaling behavior: (i) high expression in normal tissue types and strong changes in cancer, (ii) high expression, but small changes, and (iii) low expression with small changes.

4/f Suggesting novel drug targets

We analyzed the drug target relevance of human signaling proteins by examining 4 key properties: disease-relatedness, localization in the plasma membrane, enzymatic functions, and kinase domain content. To identify the most promising drug target candidates from the following two sets: (i) list of novel signaling proteins, (ii) human multi-pathway proteins and proteins participating in cancer-related cross-talks. Some of these proteins could be specific and proper targets, some of them could be too central and aspecific.

Analysis of signalogs predicted signaling pathway memberships for 5 currently used drug target proteins and suggested 14 additional proteins that can be used as novel drug target candidates.. Our predictions may (i) reveal novel therapeutic intervention

points (e.g., the use of signalogs as novel targets to block specific pathways); (ii) suggest novel applications of current drugs to diseases, where the newly predicted signaling pathway of their target is relevant, and (iii) help to identify possible side effects of currently used drugs.

After listing the properties of multi-pathway proteins and proteins participating in cancer-related cross-talks relevant for drug target selection, we suggested 4 novel drug target candidates. One of them, ROR2, was recently proposed as a novel chemotherapeutic target, while the other 3 are known to be non-specifically affected by anti-inflammatory drugs.

Summary:

1. We developed a signaling pathway database, called SignaLink. Its uniform curation rules and data structure allow the system-level examination of the signaling network. SignaLink contains the signaling pathways of 3 metazoans, the nematode *C. elegans*, the fruit fly *D. melanogaster*, and humans.

2. We compared SignaLink with 3 existing pathway databases. SignaLink was found to be better both in quantitative and qualitative properties.

3. We created several novel methods to visualize signaling networks.

4. We identified 253 novel signaling proteins, called signalogs, and verified their novelty.

5. We compared the pathways and their cross-talks, and found that only in humans every pathway can cross-talk. We found that in humans, cross-talk expression is tissue and disease-specific, which underscores its importance in development and medicine. We could identify 3 cross-talk expression types.

6. Examination of signalogs, multi-pathway proteins, and proteins important in cross-talks of cancer cells, allowed us to create a short list of possible novel drug targets.

List of publications:

Publications directly related to the thesis

1. **Korcsmaros T** *, Farkas IJ *, Szalay MS, Rovó P, Fazekas D, Spiro Z, Böde C, Lenti K, Vellai T, Csermely P (2010) Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. *Bioinformatics* **26**:2042-2050
   IF: 4,9 **Number of independent citations:** 3

2. **Korcsmaros T ***, Szalay MS *, Rovó P, Palotai R, Fazekas D, Lenti K, Farkas I J, Csermely P, Vellai T (2011) Signalogs: orthology-based identification of novel signaling pathway components in three metazoans. *PLoS ONE* **6**(5), e19240
   IF: 4,4

Publications not directly related to the thesis

1. Nardai G, **Korcsmaros T,** Csermely P (2002) Reduction of the endoplasmic reticulum accompanies the oxidative damage of diabetes mellitus, In: *Redox regulation* (eds.: A. Pompella, G. Banhegyi and M. Wellman-Rousseau), NATO Science Series, **I/347**, 281-289

2. Nardai G, **Korcsmaros T**, Papp E, Csermely P (2003) Reduction of the endoplasmic reticulum accompanies the oxidative damage of diabetes mellitus. *Biofactors* **17**, 259-267
   IF: 1,9 **Number of independent citations:** 12

3. Papp E, **Korcsmaros T**, Nardai G, Csermely P (2004) Changes of cellular redox homeostasis and protein folding in diabetes, In: *Cellular dysfunction in atherosclerosis and diabetes - Reports from bench to bedside* (eds.: M. Simionescu, A. Sima, D. Popov), Plenum Press, 228-235

4. Nardai G, Stadler K, Papp E, **Korcsmaros T,** Jakus J, Csermely P (2005) Diabetic changes in the redox status of the microsomal protein folding machinery. *Biochemical and Biophysical Research Communications* **334**, 787-795
   IF: 3,0 **Number of independent citations:** 14

5. Nardai G, Papp E, **Korcsmaros T**, Stadler K, Jakus J, Csermely P (2005) Possible links between metabolism and oxidative protein folding. Conseqences of a diabetes study, In: *Redox regulation* (eds.: A. Pompella, G. Banhegyi and M. Wellman-Rousseau), NATO Science Series, **363**, 101-109

6. Papp E, Nardai G, Sreedhar AS, **Korcsmaros T,** Csermely P (2005) Effects of unfolded protein accumulation on the redox state of the endoplasmic reticulum, In: *Redox regulation* (eds.: A. Pompella, G. Banhegyi and M. Wellman-Rousseau), NATO Science Series, **363**, 111-119

7. Papp E, Szaraz P, **Korcsmaros T**, Csermely P (2006) Changes of endoplasmic reticulum chaperone complexes, redox state, and impaired protein disulfide reductase activity in misfolding alfa-1-antitrypsin transgenic mice. *FASEB Journal* **20** (7): 1018-20
   IF: 6,7 **Number of independent citations:** 12

8. **Korcsmaros T**, Kovacs IA, Szalay MS, Csermely P (2006) Molecular chaperones: The modular evolution of cellular networks. *Journal of Bioscience* **32** (3): 441-446
   IF: 1,0 **Number of independent citations:** 15

9. Szalay MS, Kovács IA, **Korcsmaros T**, Böde C, Csermely P (2007) Stress-induced rearrangements of cellular networks: consequences for protection and drug design. *FEBS Lett.* **581**(19):3675-80
   IF: 3,4 **Number of independent citations:** 15

10. Böde C, Kovacs IA, Szalay MS, Palotai R, **Korcsmaros T,** Csermely P (2007) Network analysis of protein dynamics. *FEBS Lett.* **581**(15):2776-82
    IF: 3,4 **Number of independent citations:** 21

11. Csermely P, **Korcsmaros T**, Sulyok K (eds., 2007) Stress Responses in Biology and Medicine: Stress of Life in Molecules, Cells, Organisms, and Psychosocial Communities. *Annals of the New York Academy of Sciences*, **1113**, pp. 366
    Number of independent citations:1

12. **Korcsmaros T**, Szalay MS, Böde C, Kovács IA, Csermely P (2007) How to design multi-target drugs: Target-search options in cellular networks. *Exp. Op. Drug Discovery* **2** (6): 799-808
    Number of independent citations: 22

13. Kovacs I, Csermely P, Korcsmaros T, Szalay MS (2007) *WO patent application* WO 2007093960
    Number of independent citations: 1

14. Csermely P, **Korcsmaros T**, Kovács IA, Szalay MS, Sõti C (2008) Systems biology of molecular chaperone networks. In: The biology of extracellular molecular chaperones. *Novartis Foundation Symposium Series* **291**, Wiley, pp. 45-58
    Number of independent citations: 5

15. Farkas IJ, **Korcsmaros T**, Kovács IA, Mihalik Á, Palotai R, Simkó GI, Szalay KZ, Szalay-Bekő M, Vellai T, Wang S, Csermely P (2011). Network-based tools in the identification of novel drug-targets. *Sci. Signal.* **4**, pt3

Summary

Scientific publications:
11 articles, 4 book chapters, 1 book editing, 1 patent application
Cummulative imapct factors: **28,7** Number of independent citations: **132** (based on ISI, Scopus, and Google Scholar; as of April 2011)

Acknowledgement

- Prof. Péter Csermely
- Dr. Tibor Vellai
- Dr. Balázs Papp

- Dr. Illés Farkas
- Máté Szalay-Bekő
- Dávid Fazekas
- Petra Rovó
- Zoltán Spiró
- Lilian Zsákai
- Dr. Csaba Böde
- Robin Palotai
- Gábor Szuromi
- Dr. Katalin Lenti

- Dr. Ferenc Jordán
- Prof. Tamás Vicsek
- Prof. Gábor Vattay
- Dr. István Csabai
- Members of the Vellai-lab and NetBiol-group at the Eötvös Lorand University
- Library of the Semmelweis University
- My wife, Dia Papp and my family and other animals