



## The TRANSPATH signal transduction database: a knowledge base on signal transduction networks

Frank Schacherer<sup>1,2</sup>, Claudia Choi<sup>2</sup>, Ulrike Götze<sup>2</sup>, Mathias Krull<sup>2</sup>, Susanne Pistor<sup>2</sup> and Edgar Wingender<sup>1,2</sup>

<sup>1</sup>GBF German Research Centre for Biotechnology and <sup>2</sup>Biobase Biological Databases GmbH, Halchtersche Strasse 33, D-38304 Wolfenbüttel, Germany

Received on December 11, 2000; revised on April 10, 2001; accepted on May 24, 2001

### ABSTRACT

TRANSPATH is an information system on gene-regulatory pathways, and an extension module to the TRANSFAC database system (Wingender *et al.*, *Nucleic Acids Res.*, **28**, 316–319, 2000). It focuses on pathways involved in the regulation of transcription factors in different species, mainly human, mouse and rat. Elements of the relevant signal transduction pathways like complexes, signaling molecules, and their states are stored together with information about their interaction in an object-oriented database. The database interface provides clickable maps and automatically generated pathway cascades as additional ways to explore the data. All information is validated with references to the original publications. Also, references to other databases are provided (TRANSFAC, SWISS-PROT, EMBL, PubMed and others).

**Availability:** The database is available over (<http://transpath.gbf.de>) for interactive perusal. As an exchange format for the data, eXtensible Markup Language (XML) flatfiles and a Document Type Definition (DTD) are provided.

**Contact:** frs@biobase.de; cch@biobase.de; ulg@biobase.de; mkl@biobase.de; ewi@biobase.de; spi@biobase.de

### 1 INTRODUCTION

Cells, especially those of a complex multicellular organism, have to act and react to each other and to external influences in a well concerted manner. Thus, if we want to understand cellular behaviour and its responses to external signals, or want to influence it in a predictable manner, we have to understand the pathways through which these signals are mediated into and within the cell. Signal transduction pathways regulate the activity of many transcription factors (Montminy, 1997) and practically all oncogenes encode aberrantly functioning members of such pathways coupled to growth-regulating signals (Egan and Weinberg, 1993). Biological signaling pathways also interact with each other to form complex

networks. These networks show emergent properties like signal integration across multiple time scales or self-sustaining feedback loops which are not present in the isolated pathways (Bhalla and Iyengar, 1999).

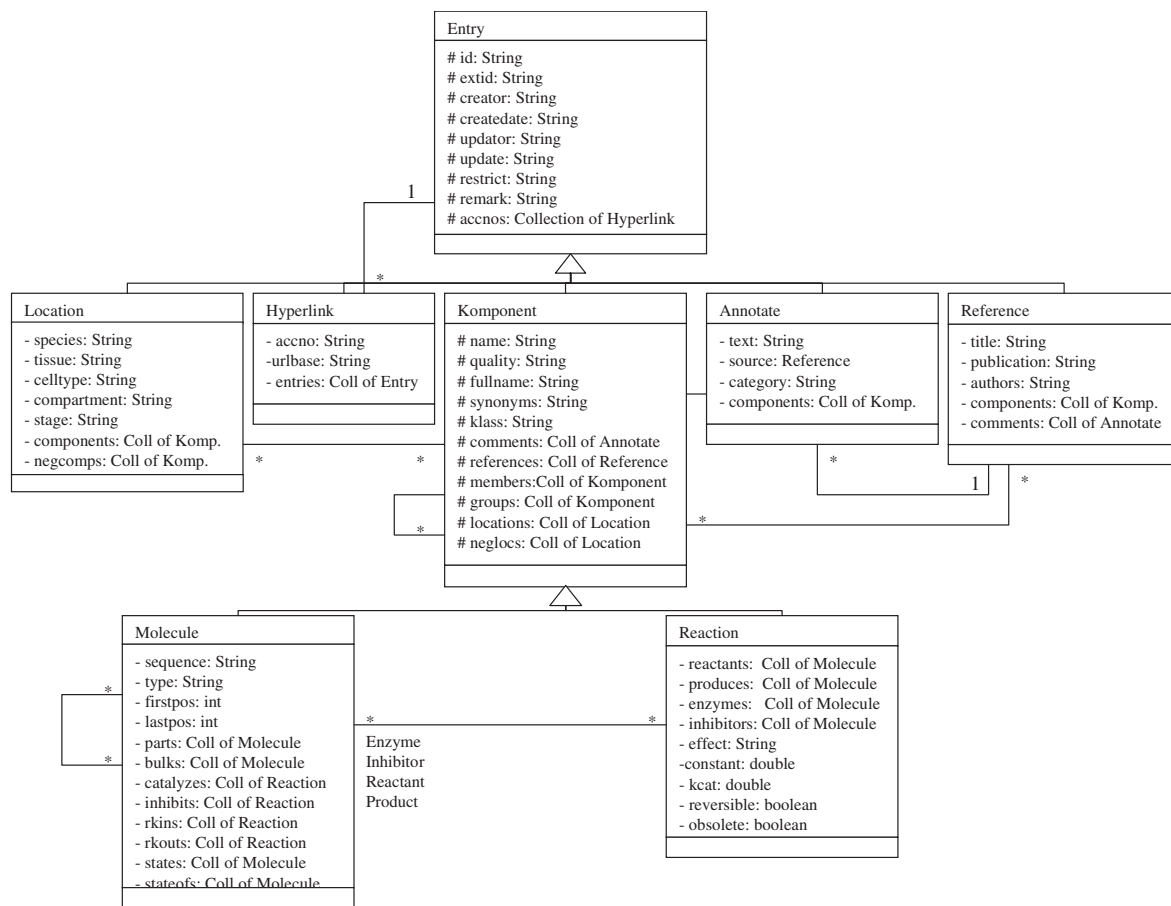
Knowledge about the principle mechanisms of signal transduction and regulation mechanisms of individual macromolecules in signaling pathways has multiplied in the last decade. Now it is growing at a rate that makes it difficult to keep up with. The huge and ever more rapidly growing amount of signal transduction data demands for a database that stores and organizes this knowledge, providing simple and fast access to the information. The complexity created by the cross-talk between pathways makes it virtually impossible to infer by hand all the consequences that follow after modification of one part of the network. To this end, computer-aided simulation will have to be used. It can only be successful on the basis of a comprehensive and detailed dataset.

### 2 METHODS AND ALGORITHMS

The database has been established under an object-oriented database management system (POE, 1999). As an interface to the database, Java and Object Query Language (Catell and Barry, 1997) are used in servlets, providing access over the www. The data are stored as a bipartite graph. To visualize the data, expanded graph traversal algorithms (Yellen and Gross, 1998) are used, which allow for searches that make use of protein family information. The object-oriented system allows to navigate the network by reachability along object references during pathway building, where in a relational system large numbers and consecutive joins would be necessary. A full class layout and detailed descriptions of all classes can be found online under <http://transpath.gbf.de/intro/tech/api>.

### 3 IMPLEMENTATION AND RESULTS

There are 19312 molecules entries and 3094 interactions in the database currently and it is updated daily. Of the molecules, 10073 were pre-imported from the



**Fig. 1.** UML class layout of the current implementation. Only one relation is shown between two classes, even if several exist. For the key relations between molecule and reaction the roles are given. To maintain clarity, methods for the classes are not shown.

public SWISS-PROT database (Bairoch and Apweiler, 2000) to avoid redundant work. Another 6109 molecule entries were generated as orthologue groups for these proteins. New molecule entries 3130 were added. They can be subcategorized as 1731 families, 881 molecules, 478 complexes and 40 motifs. All 3094 interactions are retrieved from original literature.

Figure 1 shows the class layout of the database as an UML community diagram.

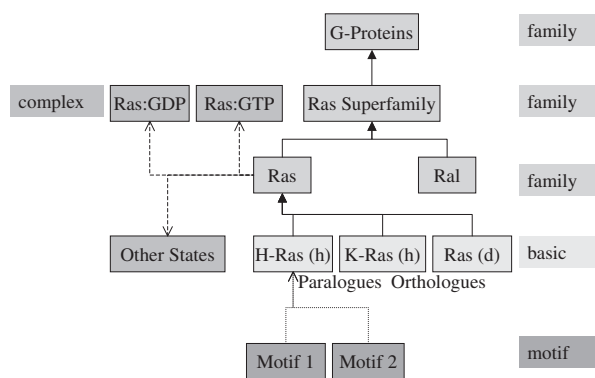
Interactions are modelled in three ways:

(1) As chemical reactions with reactants and products and, if applicable, a single enzyme and inhibitor. This class includes complex formations, phosphorylations, translocations and all other interactions for which the distinct states of the molecules involved are known. To enable the system to be used as the basis for simulation it is necessary to include rate constants in the reactions and different entries for different states of a molecule.

(2) In addition to this mechanistic view, interactions can be stored as activation and inhibition pointers providing a semantic view, which corresponds to the schematic drawings familiar from the literature, and is useful for those cases where the interaction mechanism is yet unknown or indirect (similar data is provided by the CSNDB database (Takai-Igarashi and Kaminuma, 1998)).

(3) Non-directional interactions can be used in those cases where it has only been shown that two compounds interact, but the effect of the interaction is unknown, like is the case for data from protein-interaction or binding studies (similar data is stored in the BIND and DIP databases (Bader and Hogue, 2000; Xenarios *et al.*, 2001)).

Modification, for example by covalent binding or complexation, can change a molecule's signaling behaviour. For this reason, one entry per gene or per splice variant would not be sufficient to capture the signaling behaviour.



**Fig. 2.** Hierarchical relations and roles for molecules, shown for a subset of the Ras superfamily. Basic molecules are translated proteins or small molecules that have mass. Family molecules are groups of related molecules or of molecule groups. Motifs are structural or sequence motifs of basic molecules. Complexes and other states change the availability of the molecule for reactions.

Also, motifs of a protein are often responsible for its signaling behaviour (see Hunter, 2000), and it should be possible to link the signaling reaction to the motifs instead of the whole molecule. In TRANSPATH, each molecular activity and each motif is represented by its own entry.

Grouping molecules into families is essential for a usable signal transduction database. Publications often do not exactly specify which member of a protein family was used in an experiment. When the family relationships are stated in a formal way, it is possible to capture the statements from the literature correctly, and it is easy to write algorithms that exploit them in user queries. In TRANSPATH, each family is represented by its own entry. The family classification follows the use of families in the extracted literature.

Figure 2 shows these relationships as a hierarchy.

Subcellular location plays an important role in the activity of many compounds. For example transcription factors like NF- $\kappa$ B are only active in the nucleus, Ras is only active after recruitment to the inner plasma membrane. Every data item in TRANSPATH can be associated with a location object, which also incorporates information about cell types, developmental stages, organs, and species specificity.

## 4 DISCUSSION

### 4.1 Data acquisition and updates

All interactions and molecules<sup>†</sup> are extracted manually out of the literature by academically trained annotators to ensure a high quality of the database content. External

<sup>†</sup> Except for the initially imported SWISS-PROT entries and groups.

experts are welcome to curate pathways in their area of expertise, but there will be no public interface for data submission to the database. The public version of the database will be updated yearly.

### 4.2 Comparison with other work

We extend the approach taken by CSNDB (Takai-Igarashi and Kaminuma, 1998), and offer detailed, species-specific reactions extracted from primary literature in addition to overview reactions from reviews.

TRANSPATH differs from interaction databases like BIND (Bader and Hogue, 2000) and DIP (Xenarios *et al.*, 2001) in two ways: first, in TRANSPATH no mass-generated data is stored. Second, most reactions in TRANSPATH show the direction of signal transfer, while interactions are undirected. Therefore, TRANSPATH can be used to annotate the function of such interaction data, while it in turn can benefit from the additional evidence for its reactions from those databases.

One direct advantage of directionality is that we can infer upstream or downstream events in signaling cascades, which would not be possible with undirected relations.

### 4.3 Querying the signaling network


To make a database usable, it must be possible to answer the questions the user has about the data. Natural questions to a pathway database often are more concerned with pathways and networks than with a single element, for example: what are the downstream effects of the presence or absence of some molecule? Are there crosstalking points between two pathways? Are there common regulatory elements for a set of genes?

A metaphor that allows these kind of *pathway queries* is implemented in TRANSPATH through graph traversal algorithms. It is possible to bound the search by a certain distance, to limit it to a species, and to generate shortest paths between two molecules. More elaborate queries are made possible by selecting certain sets of molecules as *filters* or *boundaries*. Filters act to highlight or select certain elements or paths that have been observed in the traversal, boundaries limit the search to a subgraph. To take into account family and state relationships in the database, the search can be extended to traverse these relations too. Pathway queries are a flexible and extensible way to request data.

### 4.4 Displaying the results

With pathway queries, we can ask the questions. To understand the answers TRANSPATH provides five different views on the data:

- (1) Tabular data that present the attributes of a single object in the database (Figure 3).
- (2) Linear paths that show an ordered list of molecules connecting two molecules of interest (Figure 4).



### Ral - Results in Transpath DB

The TRANSPATH database is free for users from non-profit organizations - It is an extension module to the TRANSFAC database on transcription factors and their binding sites.

**Build a cascade for this entry**

**Next Query**

**Result Storage**  
[STORE](#) | [ADD](#)  
[VIEW](#) | [CLEAR](#) | [EDIT](#)

**small G-protein**  
 Literature References: [Boguski M. S., McCormick F.](#)  
 Feeds Reaction: [small G-protein -> PLD](#)  
 Fed by Reaction: [GPCR -> small G-protein](#)

**Ras superfamily**  
 Located in: [inner plasma membrane](#)  
 Literature References: [Krauss, G., Malumbres, M., Pellicer, A., Pawson, T., Post, G. R., Heller, Brown, J.](#)  
 Comments:  
 physiological function: [promotes both cell death and cell survival through interactions with distinct effector proteins](#)

## Ral

ID: M0000000063  
 Created by: ffs  
 Date of Creation: 23.06.1999 10:59:07  
 Date of last modification: 11.05.2000 13:51:33  
 Name: Ral  
 Member of Group or Family:

- [Ras superfamily](#)

**Subtypes:**

- [RalA](#)
- [RalB](#)

**Literature References:**

- [Malumbres, M., Pellicer, A.](#)

**Feeds Reaction:**

- [Ral -> PLD](#)
- [Ral -> RBP1](#)

**Fed by Reaction:**

- [RalGDS -> Ral](#)

Your request took 0 seconds to answer.

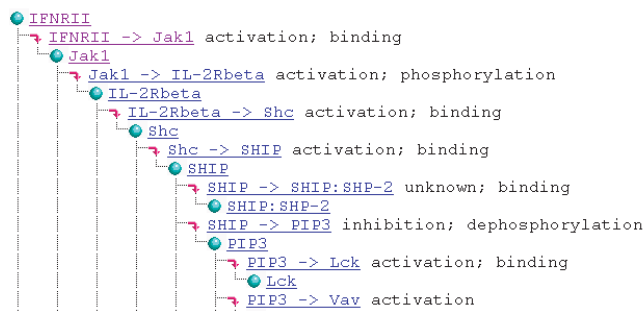
**Fig. 3.** Example output for a single molecule entry. In this example, the option to show family information was selected, printing abstracts of families this molecule belongs to above it.

[IL-1alpha](#), [IL-1RI](#), [RhoA](#), [JNK](#), [c-Jun](#), [AP-1](#)

**Fig. 4.** Example output for a path between IL-1alpha and AP-1. The connecting interactions can also be included if the user so wishes.

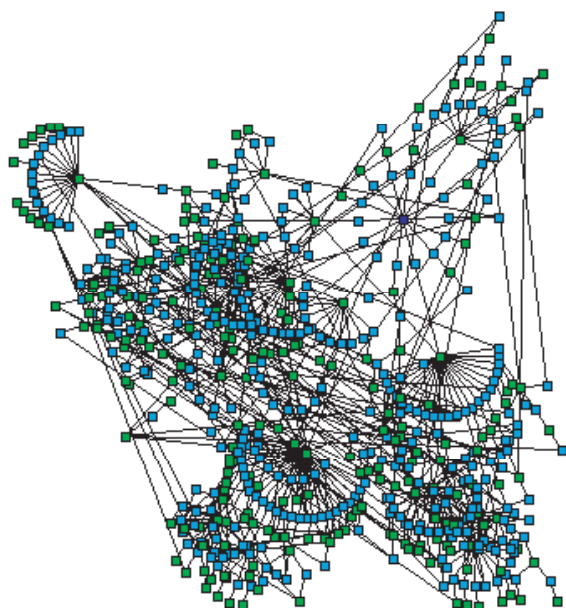
- (3) Tree-like cascades that show which elements are downstream or upstream of a core component (Figure 5).
- (4) Graph layouts that show the full network as connections between nodes, according to its topology (Figure 6).
- (5) Hand drawn, clickable maps that enrich the topological layout with additional information, like subcellular location or molecule type (Figure 7).

TRANSPATH provides a knowledge base which goes beyond the approach of traditional gene or sequence databases by focusing on the interactions between the stored data items. By building up the signaling network from single interactions instead of using predefined pathways, it becomes possible to explore the pathways through the graph in an unbiased way.



**Fig. 5.** Example output for a downstream cascade. The start for the pathways downstream of IFNR1I are shown. To limit the size of the image, it has been cut off after the first few steps.

The next step will be to analyze the reaction graph and infer general properties of the signal transduction pathways involved. With more and detailed data, it might also become feasible to run simulations to obtain suggestions for the response behaviour of such networks to extracellular signals, which then may be used to drive experimental research.



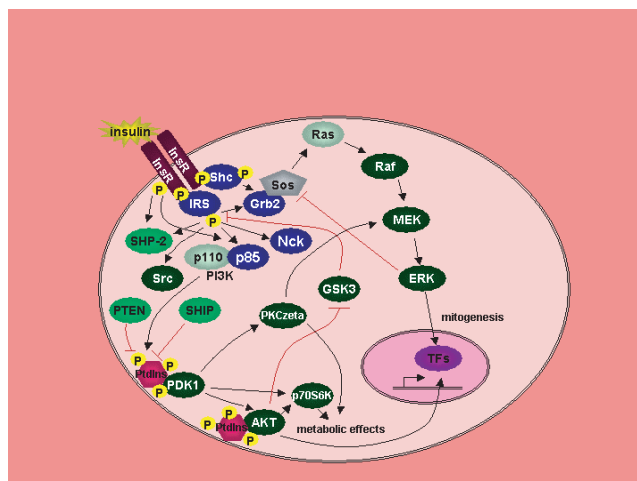
**Fig. 6.** Network around the ‘Ras’ entry, the square in the circle on the upper right, as displayed in the Otter (2000) tool, for which input files can be generated. Otter has to be downloaded and installed separately. The names for nodes can be shown for the node under the mouse pointer or for all nodes. One can directly make out important players associated with Ras, as they are linked to a large number of nodes: the half circle to the right above Ras centers on Raf, the large one to the right below Ras on PI3K, and the one to the left below Ras on Grb2. The largest one, in the middle below, centers on NF- $\kappa$ B.

## ACKNOWLEDGEMENTS

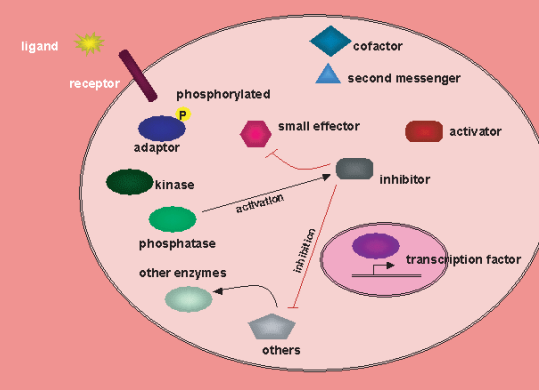
This work has been supported by a grant of the German Ministry of Education, Science, Research and Technology (BMBF; 01 KW 9629/7). The complete data set of CSNDB was generously provided by T.Takai-Igarashi.

## REFERENCES

- Bader,G.D. and Hogue,C.W. (2000) BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics*, **16**, 465–477.
- Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement trEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Bhalla,U.S. and Iyengar,R. (1999) Emergent properties of networks of biological signaling pathways. *Science*, **283**, 381–387.
- Catell,R.G.G. and Barry,D.K. (eds) (1997) *The Object Database Standard: ODMG 2.0*. Morgan Kaufmann, San Francisco, CA.
- Egan,E.S. and Weinberg,R.A. (1993) The pathway to signal achievement. *Nature*, **365**, 781–783.
- Hunter,T. (2000) Signaling—2000 and beyond. *Cell*, **100**, 113–127.
- Montminy,M. (1997) Transcriptional regulation by cyclic AMP. *Annu. Rev. Biochem.*, **66**, 807–822.
- Otter (2000) Otter. <http://www.caida.org/tools/visualization/otter/>.



### Legend



**Fig. 7.** Example image map and legend. Maps like this one for insulin signaling provide a first overview over the important molecules in a pathway, and are linked to the more detailed information in the database. Not all elements from the corresponding network in the database can be included in these overviews. Element colors and forms are keyed to classes of molecules. Currently 15 such maps are provided.

Otter is a historical CAIDA tool used for visualizing arbitrary network data.

POE (1999) POET object server & Java SDK 5.1 edition.

Takai-Igarashi, and Kaminuma,T. (1998) A pathway finding system for the cell signaling networks database. *Silico Biol.*, **1**, 0012, <http://www.bioinfo.de/isb/1998/01/0012>.

Wingender,E. *et al.* (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.

Xenarios,I., Fernandez,E., Salwinski,L., Duan,X.J., Thompson,M.J., Marcotte,E.M. and Eisenberg,D. (2001) Dip: the database on interacting proteins: 2001 update. *Nucleic Acids Res.*, **29**, 239–241.

Yellen,J. and Gross,J.L. (1998) *Graph Theory & its Applications*. CRC Press, Boca Raton, FL.