Consistent re-modeling of signaling pathways and its implementation in the TRANSPATH database

Claudia Choi[1] cch@biobase.de
Torsten Crass[2] torsten.crass@med.uni-goettingen.de
Alexander Kel[1] ake@biobase.de
Olga Kel-Margoulis oke@biobase.de
Mathias Krull[1] mkl@biobase.de
Susanne Pistor[1] spi@biobase.de
Anatolij Potapov[1,2] anatolij.potapov@med.uni-goettingen.de
Nico Voss[1] nvo@biobase.de
Edgar Wingender[1,2] ewi@biobase.de

[1] BIOBASE GmbH, Halchtersche Str.33, D-38304 Wolfenbüttel, Germany
[2] Dept. of Bioinformatics, University of Göttingen, Medical School, Goldschmidtstr. 1, D-37077 Göttingen, Germany

Abstract

The data model of the signaling pathways database TRANSPATH has been re-engineered to a three-layer model comprising experimental evidences and summarized pathway information, both in a mechanistically detailed manner, and a "semantic" projection for the abstract overview. Each molecule is described in the context of a certain reaction in the multidimensional space of posttranslational modification, molecular family relationships, and the biological species of its origin. The new model makes the data better suitable for reconstructing signaling pathways and networks and mapping expression data, for instance from microarray experiments, onto regulatory networks.

Keywords: signaling pathways, signal transduction, pathway modeling, TRANSPATH database

1 Introduction

The way how an extracellular signal is transduced inside a cell to reach its target within a specific compartment has been a matter of intense biochemical research since decades. Recent high-throughput approaches such as microarray gene expression or proteomics studies gave an additional incentive to these activities and allowed to systematically connect them with genomic features for those organisms of which whole genome information is already available. This enabled and required appropriate formalization and handling of the corresponding data for which a consistent formal database representation was to be developed at first. This was done with the signal transduction pathway databases CSNDB [15],[13], aMAZE [16], TRANSPATH[1] [12],[9],[2], or PATIKA [3].

Besides, information about protein interaction networks has been gathered in a large scale [1],[11]. It is easier to gain this kind of data by high-throughput approaches than those about signaling networks, since signaling data require a minimum of functional information about the participating molecules to impose a direction of the signal to flow along the protein interaction chain, and also need to involve non-proteinaceous components along the signal flow. In any case, the biochemical mechanism by which the signal is processed is not represented in the pure protein interaction databases. It is given in some of the signal transduction databases, whereas some others restrict to the representation of the key molecules which actively transfer the signal. Both of these representation modes have their justifications and merits: The mechanistic view is

[1] TRANSPATH is a registered trademark (®) of BIOBASE GmbH, Wolfenbüttel, Germany

desirable since it allows the generation of more detailed hypotheses/predictions; however, it is frequently too confusing for human perception. Furthermore, the available data frequently do not allow modeling in mechanistic detail, hence a semantic representation seems to be mandatory anyway. For these reasons, we decided early to implement the "mechanistic" (or biochemical) view as well as the "semantic" (or molecular) view in the TRANSPATH database [12] ("biochemist's" or "cell biologist's" view in [6]). However, in addition to requiring an extra effort of manual annotation, this way of maintaining the database is prone to produce inconsistencies. We therefore revisited the data structure of the TRANSPATH database, attempting now to model pathways mechanistically as detailed as possible, deducing the semantic representation as automatically as possible, and re-sorting the different dimensions of the molecular systematics, i. e. the molecular hierarchy, molecular modifications, and the assignment to biological taxa.

2 Methods and Results

2.1 The pathway model

One of the fundamental features of signaling pathways and networks is that they are directed, principally discriminating them from mere protein interaction networks. The directionality is defined by those steps that are irreversible under cellular conditions, which are most of the catalyzed reactions in signal transduction, in contrast to interactions or complex formation reactions.

Therefore, we designed our data model as a bipartite directed graph, the node classes representing "molecules" and "reactions", the edges representing the capabilities of molecules to enter into a or exit from a reaction [12]. In our data model, we attempt to represent all steps of a pathway with all known details of the underlying reaction mechanisms. The corresponding view of these data is thus called "mechanistic". An example for such a mechanistic representation is shown for the TGF-β pathway in Fig. 1 (see also [8] for further details).

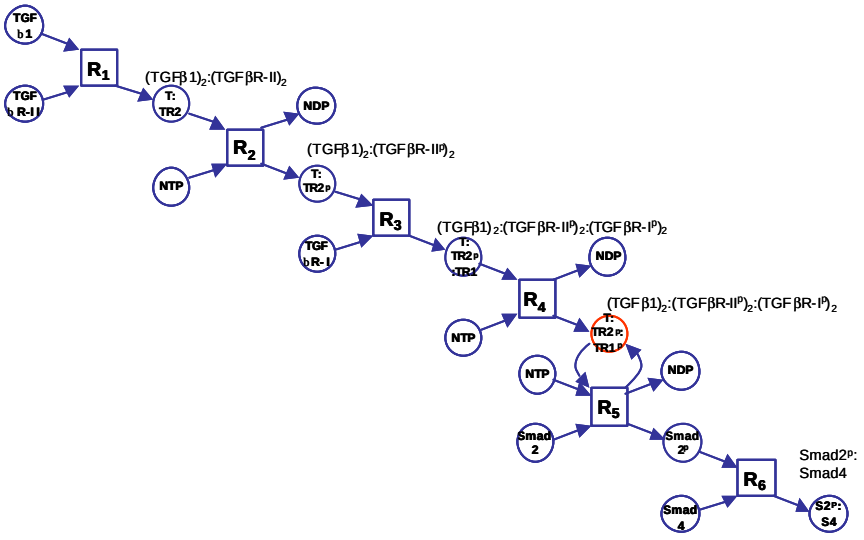

Figure 1: Bipartite directed graph modeling of the TGF-β pathway from ligand-receptor interaction to activation of a Smad transcription factor heterodimer. TGF-βR, TGF-β receptor subunits; NTP(NDP), nucleoside tri(di)phosphate; R_i, reactions along the pathway.

When doing so, it is not necessary to (manually) annotate classes of reactions, such as "complex formation" or "phosphorylation". This should (and actually does) become obvious from the type of the molecules that are consumed and produced by a reaction. Also, the fact that a reaction is catalyzed and which of the participating molecules is the catalyst should be revealed by the topology of the corresponding step, since we strictly adhere to the definition of "catalyst" as a substance that enters a reaction to reduce its activation energy and leaves it unalteredly. Such a molecule would be connected to the reaction it catalyzes by both an incoming and an outgoing edge.

In this mechanistic representation of reactions, we have a strict and consistent logical AND connection

between different incoming (or outgoing) edges of one Reaction node, since all molecules entering a reaction are needed to let it happen, and all products have to appear as its result. On the other hand, the logical connection between several incoming (or several outgoing) edges at a Molecule node is of logical XOR connection, since an individual molecule can be produced or consumed by only one particular reaction, whereas a population of molecules which is normally represented by chemical reaction equations can have a mixed origin or destination, rendering it an OR relation.

Thus, the function of an inhibitor may also be deduced from the network topology (Fig. 2). An enzyme catalyzing a reaction may be a reactant of another reaction at the same time. Because of the XOR relation of both edges that link the enzyme (molecule) node with two reactions, it follows that both of them are "competitive" to each other. If one of these reactions is consuming the enzyme, i. e. having a directed edge from the enzyme molecule to that reaction but no edge pointing back, and/or no catalytic activity is indicated for the product of that reaction anymore, this reaction will be inhibitory to the enzyme function.

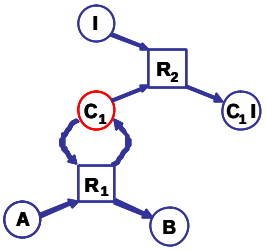

Figure 2: Typical network topology of an inhibition. C symbolizes a catalyst, I an inhibitor, A and B other participating molecules, R_1 and R_2 represent the reactions.

Interactions (mostly: protein-protein interactions) are modeled as complex formations. These are reversible reactions, which are dissolved into a forward and a back reaction (complex formation and complex dissociation). However, inside complexes, many highly relevant reactions occur that should be represented in the model (and the corresponding database) in a way that makes them amenable to the same formalism to deduce molecular and reaction functions, but also to make them distinguishable from reactions occurring amongst the same components freely in solution.

To enable this, we model the formation of a complex as one reaction in the pathway. The next reaction is an aggregation of the reactions inside the complex, resulting, e. g., in the dissociation of its components and/or release of one of them in a phosphorylated state (Fig. 3). The decomposition/expansion of such an aggregated reaction may then display the reaction among the key players within the complex, and from this expansion it will become clear which roles they play in these reactions.

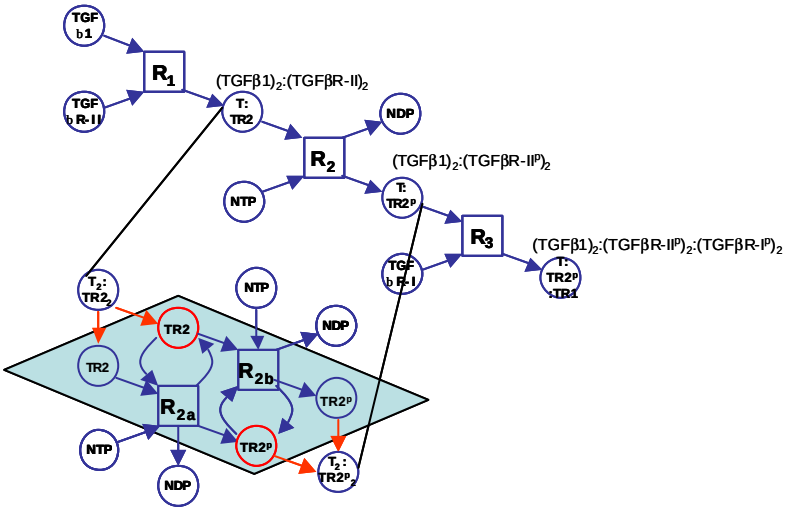

Figure 3: Modeling of processes inside a complex (shaded area) by virtual decomposition (red arrows). Cross-phosphorylation of TGFβR-II subunits is as suggested by Luo & Lodish [10].

2.2 Three-layer data model

The Evidence Level: The primary source of the information contained in the database are individual scientific publications. However, each publication usually deals with a limited number of reactions, and frequently only one interaction or reaction has been investigated in detail, giving all experimental evidence for its proof. This kind of information is essential to assess the quality of a reconstructed pathway, and to facilitate this, an extended quality assessment system which numerically codifies the reliability of the methods employed and the "authenticity" of the materials used.

The Pathway Level: However, these experimentally evidenced reactions, as close as they are to the originally published data, are extremely heterogeneous and redundant: They largely differ in the experimental depth and reliability, thus providing distinct views on the same reaction which can be aggregated. The example given in Fig. 1 is for this level.

Since it is very hard, if not impossible, to computationally identify those reactions that provide redundant information from the Evidence level and therefore can and should be aggregated, this is subject to manual annotation by experts. During this curation process, the experts gather all experimentally evidenced reactions and condense them into one mechanistically modeled reaction. These reactions can then be used to reconstruct non-redundant pathways and networks. Therefore, they constitute an own reaction level in the database, the "Pathway Level", consisting of individual pathway steps, which abstracts from (though it is linked to) the individual entries on the "Experimental Evidence Level".

The Semantic Projection Level: To get a quick overview, biologists are used to work with simplified schemes focusing only on those components along a pathway that are directly involved in processing the signal. This kind of representation is frequently used in the graphical schemes of review articles. They do not exhibit all reaction partners of the individual steps, e. g. generally omitting small abundant molecules such as H_20, CO_2, ATP, etc., but also some auxiliary molecules (adapters etc.) are often omitted. Also, modified forms and complexes are normally abstracted, and rather the basic state of the active component is shown. But there may be exceptions, depending on what the individual author wants to emphasize, thus opening up some subjectivity in this kind of representation.

Such semantic reactions were also manually annotated in the TRANSPATH database. In future, they will be mainly automatically derived from the underlying steps on the Pathway level. For this, a set of rules is currently being established for how the entities that shall appear in the semantic representation are to be deduced from the mechanistically stored network. These rules should approach the intuitive expectation and understanding of biologists, but by applying computer-manageable rules, provide an unbiased projection of the detailed layer underneath.

Moreover, semantic reactions need specific annotation about its type (e. g. phosphorylation, interaction, proteolysis, etc.), and its effect (activation, inhibition). As for the reaction type, a hierarchically organized system (an "ontology") of signaling reactions has been established. The different reaction types can be read out from the underlying mechanistic representation on the Pathway level, and the role of the individual components (substrate, cofactor, catalyst, etc.) can be assigned by the system. As for the reaction effect, the situation is somewhat more complicated since the effect always refers to a subsequent step and is not a property of the reaction *per se*.

2.3 Reaction relations / chains

From what has been said before, it becomes obvious that we need another specific kind of information that exceeds the representation of a single pathway step: the relation between two reactions. This is relevant for the semantic projection level only, and should be amenable to an automatic read-out from the pathway level.

As for two subsequent reactions, the preceding one may have an activating or an inhibitory effect on the subsequent one, and even opposite effects on two different subsequent reactions. This can be deduced from the underlying mechanistically complete representation, if the effect is an all-or-nothing one. For instance, if one molecule state is not able to enter a certain reaction, whereas another one (e.g., after phosphorylation or interaction with another protein) can do, this is clearly seen on the pathway level. In this case, the phosphorylation reaction would exert an activating effect on the molecule that enters the next step. The local pathway topology of inhibition has already been discussed above (Fig. 2). However, if there are only quantitative differences, this would require additional information which is presently not appropriately modeled (see below, Discussion). As a result, the semantic representation of the pathway shown in Fig. 1,

still modeled as a bipartite directed graph, would look as shown in Fig. 4. Note that there is no 1:1 correlation of reactions in the mechanistic and the semantic representation. Annotation of the reaction types would have to assign "phosphorylation" to R_4, and "interaction" to R_5, for instance. $R_{1/2}$, however, has to include the ligand-receptor "interaction" as well as the TGFβ-RII subunit "autophosphorylation". The reaction "effect" would be "activation" throughout since each of these reactions enables the next one.

Figure 4: "Semantic projection" of the TGF-β signaling path.
The reaction numbers correspond to those given in Fig. 1.

Consequently, these reaction effects are now given in a separate table, Reaction Relations. In addition to these Reaction relations, we need a related, nevertheless different type of connector between reactions, "Chains", which copes with our incomplete knowledge about pathways and their locations. The general maps illustrating the architecture of a signaling network are usually reconstructed for an abstract cell. Similar to the metabolic "master pathway" maps of KEGG [6], they do not consider the occurrence of individual molecule nodes in specific cell types, tissues, or organs. However, this kind of information is principally available, can be entered into the database and can be used as a filter for pathway reconstruction. The same is true for intracellular locations, as far as it refers to well-characterized cellular compartments. However, there are examples that signaling components agglomerate somewhere in the cytoplasm, in an ill-defined structure, thus providing specificity to the signal transduction which is hard to predict just from the components involved. As a result, predictions of possible signaling networks may comprise numerous false-positive cross-connections, mainly due to "irrelevant" pathway cross-talks [4] and leading to pathways which have never been experimentally proven. This way of generating hypotheses may be a wanted effect for some users, but is confusing for many others.

As a solution for this problem, those reactions which have been experimentally shown to follow each other in a certain experimental system, are clamped together as a "Chain". These chains can be given priority when building pathways or networks, but this is optional and left up to the user.

Chains can be modeled for series of semantic or mechanistic reactions, or even mixed ones, and can even include "indirect" reactions. Indirect reactions, which are always semantic ones, are biochemical or genetic effects of a certain molecule on a remote node (another molecule or a gene), without detailed knowledge about the molecules and reactions in between. In an extreme case, a set of indirect reactions can tell us about the genes induced by a certain extracellular ligand, without any knowledge of the pathway(s) mediating this effect. Connecting indirect reactions with additional pathway information can prove useful since (i) this adds, in general, more knowledge to the effects a certain signaling molecule exerts on a given system, (ii) this may provide additional, though somewhat circumstantial, evidence for a reconstructed pathway that connects the same components, and (iii) it may in turn provide a hypothesis for the gapped knowledge of this indirect reaction.

2.4 The hierarchies

Each molecule entering or leaving a reaction is to be characterized by a number of attributes: the biological species, the tissue, cellular or intracellular location, its molecular/functional classification, its posttranslational modification status, and its complex status. Even when neglecting the latter for the time being since it has some particular properties, the remaining characteristics span a multidimensional space in which the coordinate vector determines the overall status of each molecule. This is complicated by the fact that the description of each dimension has to be hierarchically organized to cope with incomplete knowledge so that whenever all detailed information is not available, more generic items can be used instead.

Posttranslational modification:
To start with the last dimension, the posttranslational modification: At the first glance it seems reasonable to establish a hierarchy starting from the unmodified molecule, with the kind of modification underneath (phosphorylation, acetylation, methylation, ubiquitination, etc.), followed by the residue and position of the modification. However, we have to deal with two kinds of problems: Combinatory variety of modifications,

and incomplete knowledge. For the first issue, we have to take into account that each modification may occur at multiple positions, alone or in combination, and that different modifications also may occur simultaneously at different positions. This has to be solved by describing each modification at each position separately, then defining each molecular species by the appropriate combination of modifications (Tab. 1).

Doing so, we have to differentiate between the molecule in its knowingly unmodified state or the molecule in an unknown, undefined state, or the molecule in general, in any state. Thus, any behavior of the molecule that is known to refer to it independently of its modification has to be linked to the molecule class entry *M*. A specific behavior of the molecule which may refer to a specific modification state, including the state "unmodified", and may depend on it is linked to the "undefined" state of *M*, *M**. Thus, whenever it is clear that a property of the molecule is dependent on its modification state though the exact modification required to exert this property is not known, the feature (attribute) has to be linked to the state "undefined". Assigning a property to *M** precludes any defined modification state.

Different from this state, the distinct modification have to be indicated, including "unmodified" (M^0), phosphorylated (M^p), acetylated (M^a) etc. Assigning M^0 automatically excludes any of the other modified states, whereas all other modification states are principally compatible with each other. Within each of these states, we have to differentiate between the knowledge that a certain modification is required for a certain behavior, but the exact location is unknown (modification location "undefined", from the cases where this modification can be exactly located (M^{Xnm}, with X, the amino acid residue, in position *n* subject to the modification of type *m*). Setting the location "undefined" for a certain modification precludes any location definition for this modification, but allows location definitions for each other type of modification. The descriptions of different modification locations can be arbitrarily combined.

In this way, it is possible to clearly differentiate between "unmodified" (M^0), "modification not known" (*M**), and "modification does not matter" (*M*) descriptors.

Table 1: Hierarchy of molecule modifications

Polypeptide	**Modification**	**Position**	**Status 1**	**Status 2**	**Status 3**	**Status 4**	**Status 5**	**Status 6**
Molecule *M*	undefined		+					
	unmodified			+				
	phosphorylated	undefined			+			
		Pos1 (Y)				+		
		Pos2 (S)					+	
		Pos3 (T)						+
		Pos4 (S)					+	
	acetylated	undefined						
		Pos n (X)						+
	methylated	undefined						
		Pos n						
	ubiquitinylated	undefined						
		Pos n						
	...							
Symbol:			M^*	M^0	M^p	M^{Y1p}	$M^{S2p,S4p}$ $M^{(S2,S4)p}$	$M^{T3p,Xna}$

Molecular / functional classification:

Signal transduction cascades proceed through a series of steps, the number and kind of them largely varying between different signaling pathways. However, some common features can be identified: Thus, incoming signal molecules generally interact with a receptor at the beginning of each cascade, and at the end of each pathway a number of target molecules are affected. These target molecules may be defined according to the large groups of cellular effects a signal transduction pathways may be aim at (i) gene regulation events through transcription factors, (ii) metabolic events through metabolic enzymes, (iii) morphological alterations through structural proteins, (iv) secretory processes through the components of the precursor processing machineries.

To facilitate consistent modeling of the individual steps of signaling cascades, attempts were made to develop a comprehensive classification of signaling molecules. This corresponds to the branch of "regulatory

entity > regulatory component > molecular components > molecular functional components" in our previously published top-level ontology [17]. Under this, we identify a series of functionally defined signaling components such as ligands, receptors, enzymes, transcription factors, etc. (see Fig. 5A). This hierarchical classification does not exhibit a tree-like structure since many molecules have pleiotropic function, such as receptors with protein kinase activity, etc.

(A)

- adapter proteins 1
- chaperones 2
- co-factors 3
- cytoskeletal proteins 4
- enzymes 5
- GTPase-controlling signal molecules 6
- inhibitors 7
- ligands 8
- membrane-transducing components 9
- transcription factors 10
- transporters 11
- unclassified

(B)

- Family
  - Subfamily
    - Genus
      - Species

(C)

- MAPK
  - ERK
    - ERK1
    - ERK2
    - ERK3
    - ERK4
    - ERK5
  - JNK
    - JNK1
      - JNK1α1
      - JNK1α2
      - JNK1β1
      - JNK1β2
    - JNK2
      - JNK2α1
      - JNK2α2
      - JNK2β1
      - JNK2β2
    - JNK3
      - JNK3α1
      - JNK3α2
  - p38
    - p38α
      - CSBP2
      - CSBP1
      - Mxi2
      - Exip
    - p38β
      - p38β1 (SAPK2b)
      - p38β2 (SAPK2)
    - p38γ
    - p38δ (ERK6)

Figure 5: Molecular / functional hierarchy of signaling components. (A) Upper levels of the hierarchy of "molecular functional components" of the previously published top-level ontology of molecular regulation in biological systems [17]. (B) Categories at the lower end of the molecule hierarchy; "genus" refers to all proteins encoded by one gene, "species" to the individual polypeptides. (C) Example for the low-level part of "MAP-kinases, MAPK". Its complete path is: enzymes > protein kinases > protein serine/threonine kinases > MAPK. "Species" in light gray indicate molecules the existence of which is known but which are not yet in the database because of lack of functional information.

At the bottom end of this classification, molecules with very similar properties (which may be paralogs), and splice variants are given. The second last level at the end of this tree-structured hierarchy is represented by genes encoding signaling components, whereas the last level gives splice variants encoded by these genes. This is exemplified with the superfamily of MAP kinases, MAPK (Fig. 1B). They may be classified into ERK, JNK and p38 families according to overall sequence similarities and (to some extent) substrate specificity. Each of these families comprise all the corresponding gene products (polypeptide "species") being grouped according to their genes ("genera").

When applying our mechanistic data representation described above, each specific molecule participating in a signaling reaction should be ideally assigned to an entry that represents a specific polypeptide (or its associations, see below). However, and as in the case of posttranslational modifications, we have to cope with the problem of incomplete knowledge again. For instance, some sources may give only vague "family" information for a certain node in the network (e. g., "JNK"). This kind of information is frequently found in review papers, and sometimes is even poorly evidence, but may be sufficiently interesting to be taken up in the database since they may allow consistent modeling of pathways which otherwise would be fragmented. Or this information was retrieved by more circumstantial evidence, concluded from the observed general properties of the corresponding component. In other cases, a reported experiment may have identified the "genus" (like JNK2) being involved in a certain reaction, but did not specify the exact splice variant. We therefore need a mechanism to link reaction to these more generic molecule entries.

Linking reactions to a generic entry always states that it is not known which molecular "species" underneath is actually involved. If all "species" of one "genus" are known to act in this way, this is indicated for each "species" separately. This is important since it must be avoided that a newly discovered "species" with different properties erroneously "inherits" a certain behavior from the corresponding parent node.

An addition to the groups of molecular components defined by this classification, there is the possibility to compose free groups whenever it has been shown that only some members of a "taxon" (class, family, etc.) share a certain property or behavior, whereas others do not. This feature even allows combining members of completely different taxa into one group.

Species hierarchy:
Many experimental systems are heterogeneous in that the gene encoding a certain component has been artificially introduced into a new cellular environment, and both the gene and the cell may well belong to different species. This information is frequently not given in the publication and can only be retrieved by going back through several layers of cited reports, if at all. As a consequence, pathway information has to deal with incomplete knowledge about the biological species as well.

To deal with this kind of incomplete knowledge, we had introduced early the level of "ortholog" entries, which abstracted information from the underlying species-specific molecules, "basic" molecule entries which were assigned to specific polypeptides with concrete molecular weights and amino acid sequences. Throughout years of annotation work, we noticed that this concept fell short in that it

- did not allow to include species-specific molecule entries on higher levels than molecular "species" as defined in the molecular classification (see above); e.g., it was not conform with the system to create a "human JNK1" entry on top of the human JNK1 splice variants;
- did not allow to restrict the range of biological species; i. e. what was described on the "ortholog" entry level always comprised all orthologous molecules from principally all species represented in the database. However, it turned out that it was possible, reasonable and even necessary in many cases to restrict this to, e. g. mammals or vertebrates.

To tackle these problems, we introduced a hierarchy of biological species which is handled in an analogous manner as described above for the molecule and modification hierarchies. In practice, previous "ortholog" entries are now mostly resolved into mammalian, vertebrate or metazoan entries.

2.5 TRANSPATH implementation

In practice, the reverse way is used during practical annotation: Having generated all required Molecule entries (see below, The Molecule hierarchy), they can be connected by one Reaction entry, defining their roles and the reaction type. E. g., the reaction "kinase A uses ATP to phosphorylate protein B at serine-*n*" is transformed into the reaction

B + ATP + A --> B(Ser-*n*) + ADP + A

Note that catalysts, by definition, enter a reaction and leave it unalteredly, therefore have to appear identically on the left and right side of the reaction equation (which actually is not an equation at all).

This is projected into the semantic reaction:

$$A \rightarrow B$$

which just says that A is forwarding the signal to B. The arrow $\rightarrow$ (or, in ASCII: -->) of this semantic $\dashv$ reaction also suggests that A "activates" B since inhibitions are usually symbolized by the blunt-ended arrow (ASCII spelling: --/).

TRANSPATH® release 5.2 contains 17,751 molecule entries (proteins and other components that transduce extracellular signals to target genes), 4,776 gene entries (information on target genes and gene expression as starting points for regulatory pathways or feedback loops), 21,582 reaction entries (gives information on interactions between signaling molecules that constitute regulatory pathways and networks), 289 pathways (canonical signaling pathways and the reaction chains they consist of) and 7,079 reference-entries linked to PubMed. A more recent integration of TRANSPATH with NetPro™ (Molecular Connections Ltd, Bangalore, India) increased the content to a total of 25,423 molecules, 55,727 reactions, 6309 genes and 17,071 references.

From individual reactions, on any of the three levels, the PathwayBuilder™ module of TRANSPATH can re-construct whole pathways by analyzing which further reactions the resulting molecules of a previous one may enter. By default, this module starts on the Pathway Level and may use the molecule hierarchy for bridging gaps of incomplete knowledge.

A second module, the ArrayAnalyer™, maps list of genes that may come from microarray gene expression studies, onto the whole signaling network stored in the database. The algorithms of both tools are based on standard shortest-path-algorithms (Dijkstra, Floyd etc.).

Moreover, in addition to modeling the signaling networks as bipartite directed graphs, TRANSPATH assigns weights to the edges that are defined by the experimental reliability of the individual reactions, and by the belonging to experimentally proven Chains (see above). A "rigidity parameter" allows the user to select the proper balance between rigidity and sensitivity during the search.

The data model described under 2.1-2.3 has been implemented since TRANSPATH version 5.1, whereas most of the features connected with the hierarchies (modification, molecule and species hierarchies) will be implemented in one of the upcoming two TRANSPATH releases (5.3 and 5.4).

A public version of TRANSPATH 5.1 including PathwayBuilder™ is freely available to users from non-profit organizations under http://www.gene-regulation.com [19].

3 Discussion

With the development and implementation of the new data model, the TRANSPATH database is providing a comprehensive information resource on signaling pathways. The combination of a couple of features makes the system unique:

- together with the NetPro™ database, the system is the largest existing resource of manually annotated signaling reactions and pathways that is freely available for academic users;
- the new data model combines the advantages of a biochemical ("mechanistic") with a molecular ("semantic") view; this alleviates the need to manually assign, e. g, "roles" to the molecules in the semantic view which rather should be derivable from the underlying mechanism;
- the expansion and decomposition of complexes, which is similar to the Compound Graph representation suggested by Fukuda and Takagi [5], allows to represent processes occurring inside these complexes by the same mechanisms that are applied for the pathway in general;
- the three-layer model allows to navigate between experimental details, an integrated mechanistically correct pathway view, and a semantic overview; providing the more abstract levels as entry points for efficient pathway building, the resulting networks will suffer much less from all kind of redundancies;
- the molecule hierarchy allows to deal efficiently with the problem of incomplete knowledge by substituting generalized information for (missing) details;
- this navigation is supported by allocating each molecule participating in a reaction in the space of posttranslational modifications, molecular relationships, and biological species assignment, thus rendering storage and usage of these data much more consistent and effective;

What needs to be done is (i) the accurate inclusion of quantitative effects, i. e. gradual differences between competing processes, (ii) to build an ontology about cells and tissues and connect individual reactions to these locations, and (iii) to build up "clean" ontologies for the individual reactions, the pathways they build up, and cellular and physiological processes these pathways trigger. The first topic is under intense study in our groups, the second has been tackled earlier by the CYTOMER database [18], which is presently under rigorous revision. The third topic may gain significant impact from the ontological developments published by Takai-Igarashi and Mizoguchi [14] as well as from the CyteWalk project recently launched at the University of Göttingen [19].

4 Acknowledgments

The authors acknowledge the numerous helpful and stimulating discussions with T. Takai-Igarashi, University of Tokyo. Part of this work was supported by the Federal Ministry of Education, Science, Research and Technology through the Bioinformatics Center "Intergenomics" (grant no. 031U210B).

References

[1] Bader, G.D., Betel, D. and Hogue, CW., BIND: the Biomolecular Interaction Network Database, *Nucleic Acids Res*., 31(1):248-250, 2003.

[2] Choi, C., Krull, M., Kel-Magoulis, O., Pistor, S., Potapov, A., Voss, N. and Wingender, E., TRANSPATH® – a high quality database on signal transduction, *Comparative Functional Genomics*, 5:163-168, 2004.

[3] Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., Nisanci, G., Cetin-Atalay, R. and Ozturk, M., PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways, *Bioinformatics*, 18(7):996-1003, 2002.

[4] Dumont, J.E., Pécasse, F. and Maenhaut, C., Crosstalk and specificity in signalling. Are we crosstalking ourselves into general confusion? *Cell. Signalling*, 13(7):457-463, 2001.

[5] Fukuda, K. and Takagi, T., Knowledge representation of signal transduction pathways, *Bioinformatics*, 17(9):829-837, 2001.

[6] Huang, S., Regulation of cellular states in mammalian cells from a genomewide view, In: *Collado-Vides, J. and Hofestädt, R. (eds.), Gene regulation and metabolism*, MIT Press, Cambridge, MA, 181-220, 2002.

[7] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M., The KEGG resource for deciphering the genome, *Nucleic Acids Res*., 32(Database issue):D277-D280, 2004.

[8] Kloos, D.-U., Choi, C. and Wingender, E., The TGF-β-Smad network: introducing bioinformatic tools, *Trends Genet*., 18(2):96-103, 2002.

[9] Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A. and Wingender, E., TRANSPATH®: an integrated database on signal transduction and a tool for array analysis, *Nucleic Acids Res*., 31(1):97-100, 2003.

[10] Luo K. and Lodish H.F., Positive and negative regulation of type II TGF-beta receptor signal transduction by autophosphorylation on multiple serine residues. *EMBO J.*, 16(8):1970-1981, 1997.

[11] Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D., The Database of Interacting Proteins: 2004 update, *Nucleic Acids Res*., 32(Database issue):D449-D451, 2004.

[12] Schacherer, F., Choi, C., Götze, U., Krull, M., Pistor, S. and Wingender, E., The TRANSPATH signal transduction database: a knowledge base on signal transduction networks, *Bioinformatics*, 17(11):1053-1057, 2001.

[13] Takai-Igarashi, T. and Kaminuma, T., A pathway finding system for the cell signaling networks database, *In Silico Biol*., 1:0012, 1998.

[14] Takai-Igarashi, T. and Mizoguchi, R., Cell signaling networks ontology, *In Silico Biol*., 4:0008, 2004.

[15] Takai-Igarashi, T., Nadaoka, Y. and Kaminuma T., A database for cell signaling networks. *J. Comput. Biol*., 5(4):747-754, 1998.

[16] van Helden, J., Naim, A., Lemer, C., Mancuso, R., Eldridge, M. and Wodak, S.J., From molecular activities and processes to biological function, *Brief. Bioinform*., 2(1):81-93, 2001.

[17] Wingender, E., TRANSFAC®, TRANSPATH® and CYTOMER® as starting points for an ontology of regulatory networks, *In Silico Biol*., 4:0006, 2003.

[18] Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüß, M., Reuter, I. and Schacherer, F., TRANSFAC®: an integrated system for gene expression regulation, *Nucleic Acids Res*., 28(1):316-319, 2000.

[19] http://www.bioinf.med.uni-goettingen.de/typo3/index.php?id=72&type=1

[20] http://www.gene-regulation.com/cgi-bin/pub/databases/transpath/search.cgi