Title: Pathway databases and tools for their exploitation: benefits, current limitations and challenges

Running Title: Pathway databases and the systems biology promise

Anna Bauer-Mehren, Laura I. Furlong†, Ferran Sanz
Research Unit on Biomedical Informatics (GRIB)‡, IMIM-Hospital del Mar, Universitat Pompeu Fabra
Barcelona Biomedical Research Park (PRBB), carrer Dr. Aiguader 88, 08003 Barcelona (Spain)

Subject categories:
- Bioinformatics
- Signal transduction

Main text:
Word count: 5,738
Characters (excluding spaces): 32,201
Characters (including spaces): 37,909

† Corresponding author. e-mail: lfurlong@imim.es, Tel.: +34 93 316 0521, Fax: +34 93 316 0550
‡ This research unit is the Biomedical Informatics Node of the Spanish Institute of Bioinformatics (www.inab.org)
Abstract
In the last years, comprehensive representations of cell signalling pathways have been developed by manual curation from literature, which requires huge effort and would benefit from information stored in databases and from automatic retrieval and integration methods. Once a reconstruction of the network of interactions is achieved, analysis of its structural features and its dynamic behaviour can take place. Mathematical modelling techniques are used to simulate the complex behaviour of cell signalling networks, which ultimately sheds light on the mechanisms leading to complex diseases or helps in the identification of drug targets. A variety of databases containing information on cell signalling pathways have been developed in conjunction with methodologies to access and analyse the data. In principle, the scenario is prepared to make profit of this information for the analysis of the dynamics of signalling pathways. But, are the knowledge repositories of signalling pathways ready to realize the systems biology promise? In this article we aim to initiate this discussion and to provide some insights on this issue.

Keywords: biological pathways / cell signalling / pathway databases / network models / systems biology

1. Introduction
The last decades of research have led to a better understanding of the processes involved in cell signalling. Cell signalling refers to the biochemical processes by which cells respond to cues in their internal or external environment (Alberts et al., 2007). With the advent of high throughput experimentation, the identification and characterization of the molecular components involved in cell signalling became possible in a systematic way. In addition, the discovery of the connections between each of these components promoted the reconstruction of the chain of reactions that subsequently gives rise to a signalling pathway. Ultimately, our ability to interpret the function and regulation of cell signalling pathways is crucial for understanding the ways in which cells respond to external cues and how they communicate with each other.

In this regard, the systematic collection of pathway information in the form of pathway databases and the application of mathematical analysis for pathway modelling are crucial. Several databases containing information on cell signalling pathways have been developed in conjunction with methodologies to access and analyse the data (Suderman & Hallett, 2007). Furthermore, mathematical modelling emerged as a solution to study the complex behaviour of networks (Alves et al., 2006; Fisher & Henzinger, 2007; Karlebach & Shamir, 2008). The models so far obtained allow formulating hypothesis that can be tested in the laboratory. Iterative cycles of prediction and experimental verification resulted in the refinement of our knowledge of cell signalling and have shed light on different aspects of cell signalling at a systems level (regulatory aspects such as feedback control circuits or architectural features such as modularity).

Furthermore, signalling cascades are not isolated units within the cell, but form part of a mesh of interconnected networks through which the signal elicited by an environmental cue can traverse (Yaffe, 2008). Ultimately, each cell is exposed to a variety of signalling cues, and the specificity of the response will be determined by the signalling mechanisms that are activated by the cue (Alberts et al., 2007). Recent research highlights the importance of the so called crosstalks between pathways, such as the recently published connections between signalling through the purinergic receptors and the Ca²⁺ sensing (Chaumont et al., 2008), the link between extracellular glycocalyx structure and nitric oxide signalling pathway (Tarbell & Ebong, 2008), the interactions between insulin and epidermal growth factor signalling (Borisov et al., 2009) and the crosstalk between phosphoinositide 3 kinase and Ras/extracellular signal-regulated kinase signalling pathways (Wang et al., 2009).

An important goal of current research is to achieve a reconstruction of the network of interactions that gives rise to a signalling pathway in a biologically consistent and meaningful manner that in turn allows the mathematical analysis of the emerging properties of the network. In this regard, comprehensive maps of signalling pathways have been developed by manual curation from literature (Calzone et al., 2008; Oda & Kitano, 2006; Oda et al., 2005). Building such reference maps requires huge effort and would benefit from information stored in databases and from automatic retrieval and integration methods. Once a reconstruction of the network of interactions is achieved, analysis of the structural features of the network and its dynamic behaviour can take place. A commonly seen architecture of signalling pathways is called “bow-tie”, in which many input and output signals are handled by a common layer constituted by a small number of conserved components. This network architecture provides robustness and flexibility to a variety of external cues due to the redundancy of reactions that are part of the input and output layers.
Robustness refers to the ability of an organism to compensate the effects of perturbations to maintain the organism’s functions (Kitano, 2007b). Such perturbations can be changes in the availability of nutrients as well as the presence of mutagens or toxins. Moreover, systems can be subjected to functional disruptions when facing perturbations for which they are not optimized, thus revealing points of fragility of the biological system (Kitano, 2007b). For instance, an undesired effect of a drug can be caused by the unwanted interaction of the drug with molecules that represent points of fragility of the physiological system (Kitano, 2007a). On the contrary, drugs can be completely ineffective when the robustness of the system compensates their action. It has been suggested that crosstalks between signalling pathways contribute to the robustness of cells against perturbations (Kitano, 2007a). In addition, the points of fragility of the system are sometimes exploited by pathogens causing diseases, or represent processes that are usually found to malfunction in particular diseases, such as cancer. Diseases that arise from dysfunction in cell signalling are usually not attributed to a single gene but to the failure of emerging control mechanisms in the network. It has been reported that the loss of negative feedback loops characterizes solid tumors (Amit et al., 2007). These diseases are difficult to diagnose and treat unless accurate understanding of the underlying principles regulating the system is in place. Thus, the interpretation of the global properties of signalling pathways has important implications for the elucidation of the mechanisms that lead to complex diseases, and also for the identification of drug targets.

Currently, there are several repositories of information on cell signalling pathways that cover a wide range of signal transduction mechanisms and include high quality data in terms of annotation and cross references to biological databases. In principle, the scenario is prepared to make use of the information for the analysis of the behaviour of the signalling pathways. Thus, are the knowledge repositories on signalling pathways ready to realize the systems biology promise? In this article, we aim to initiate this discussion and to provide some insights on this issue.

First, we present an analytical overview of current pathway databases (Section 2). In Section 3 we present the results of an evaluation exercise conducted to determine the accuracy and completeness of current pathway databases in front of an expert-curated pathway used as “gold standard”. Moreover, we propose a strategy for the use of pathway data from public databases for network modelling (Box 1). Finally, in Section 4 we discuss the strengths and limitations of the current pathway databases and their usefulness in practical biological problems and applications.

2. Pathway databases

Pathway databases serve as repositories of current knowledge on cell signalling. They present pathways in a graphical format comparable to the representation in text books, as well as in standard formats allowing exchange between different software platforms and further processing by network analysis, visualization and modelling tools. Currently, there exist a vast variety of databases containing biochemical reactions such as signalling pathways or protein-protein interactions. The Pathguide resource serves as a good overview of current pathway databases (Bader et al., 2006). More than 200 pathway repositories are listed, from which over 60 are specialized on reactions in human. However, only half of them provide pathways and reactions in computer-readable formats needed for automatic retrieval and processing, and even less support standard formats such as Biological Pathway Exchange (BioPAX) (http://www.biopax.org) and Systems Biology Markup Language (SBML) (Hucka et al., 2003).

In order to obtain a complete view of the biological process of interest, combination of information from diverse reactions and pathways is often needed. A recent publication (Adriaens et al., 2008), describes a workflow developed for gathering and curating all information on a pathway in order to obtain a broad and correct representation. However, the described process heavily relies on manual intervention. Consequently, there is a need for automation of both the pathway retrieval process and the integration of different data sources. This section is devoted to the description of main pathway databases: Reactome, KEGG, WikiPathways, PID and Pathway Commons. Table 1 lists all pathway databases and protein-protein interaction resources that are mentioned in this section.

2.1. Reactome

Reactome is currently one of the most complete and best curated pathway databases. It covers reactions for any type of biological process and organizes them in a hierarchal manner. In this hierarchy, the lower level corresponds to single reactions, whereas the upper level represents the pathway as a whole.
Reactome was firstly developed as an open source database for pathways and interactions in human. Equivalent reactions for other species are inferred from the human data (Vastrik et al., 2007) providing coverage to 22 non-human species including mouse, rat, chicken, puffer fish, worm, fly, yeast, and E. coli. Furthermore, other Reactome projects exist focusing on single species such as the Arabidopsis Reactome (http://www.arabidopsisreactome.org).

All pathway and reaction data in Reactome are extracted from biomedical experiments and literature. For this purpose, PhD-level biologists are invited to work together with the Reactome curators and editors on the curation of data on selected biological processes. Once the first outline of the biological process is created and annotated, it is inspected by peer-reviewers and potential inconsistencies and errors are fixed. Every two years the data is reviewed in order to keep it updated (Joshi-Tope et al., 2005; Matthews et al., 2009). Moreover, cross references to different databases such as UniProt (The UniProt Consortium, 2008), Ensembl (http://www.ensembl.org/index.html), NCBI (http://www.ncbi.nlm.nih.gov), Gene Ontology (GO) (Ashburner et al., 2000), Entrez Gene (Maglott et al., 2007), UCSC Genome Browser (http://genome.ucsc.edu), HapMap (http://www.hapmap.org), PubMed, as well as to other pathway databases such as KEGG (Kanehisa & Goto, 2000) are provided.

Pathways are presented as chains of chemical reactions and the same data model is used to describe reactions for any biological process such as transcription, catalysis, or binding (Matthews et al., 2007, doi:10.1038/pid.2007.3). All in all, this represents a coherent view of pathway knowledge. The data model is based on classes, such as physical entity or event. Physical entities comprise proteins, DNA, RNA, small molecules but also complexes of single entities. Proteins, RNA and DNA, for which the sequence is known, are linked to the appropriate databases. Chemical entities such as small molecules are linked to ChEBI (http://www.ebi.ac.uk/chebi/init.do). An event can be either a ReactionLikeEvent, which represents reactions that convert an input into an output, or a PathwayLikeEvent, grouping together several related events. Each class possesses properties, such as information on the type of interaction (e.g. inhibition or activation). Reactome explicitly considers the different states an entity can display in a reaction. The phosphorylated and the unphosphorylated version of a protein are, for example, represented as separate entities. Additionally, generalization is allowed. This means that if two different entities have exactly the same function in a reaction, such as isoenzymes, the reaction is only described once and the functional equivalent entities belong to the same defined set. Another interesting element of the Reactome data model is the use of candidate sets, which act as placeholders for all possible entities in a reaction, in case the exact entity involved in the reaction is not yet known.

Reactome can either be directly browsed or queried by text search using for instance UniProt accession numbers. In addition, some tools for advanced queries are provided. The PathFinder tool allows connecting an input to an output molecule or event by constructing the shortest path between both. The SkyPainter tool can be used to identify events or pathways that are statistically overrepresented for a list of genes or proteins. Moreover, Reactome data can be combined with other databases such as UniProt, by using the Reactome BioMart (http://www.biomart.org) tool.

In addition to browsing pathways via the Reactome web interface, it is possible to download the data for local visualization and analysis with other tools. Different formats are provided for pathway download, including SBML Level 2 and BioPAX Level 2 and Level 3 (for some reactions only), as well as graphical formats. Pathway files, for instance in BioPAX format, can be directly opened in Cytoscape (Shannon et al., 2003), a software for visualization and analysis of networks. Moreover, data can be programmatically accessed through a SOAP web service.

2.2. KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is not only a database for pathways but consists of 19 highly interconnected databases, containing genomic, chemical and phenotypic information (Kanehisa et al., 2008; Kanehisa & Goto, 2000). Here we concentrate on the database storing biological pathways. KEGG categorizes its pathways into metabolic processes, genetic information processing, environmental information processing including signalling pathways, cellular processes, information on human diseases and drug development. However, the best organized and most complete information can be found for metabolic pathways. KEGG is not organism specific but covers a wide range of organisms including human. The pathways are manually curated by experts using literature. In addition to the interconnection of all databases underlying KEGG, links to external databases as NCBI Entrez Gene, OMIM, UniProt and GO are provided. Pathways can either be browsed or queried by free text search. The user can search for gene names, chemical compounds or whole pathways. A tutorial on how to browse pathways in KEGG and an overview of the multiple representation formats is available (Aoki-Kinoshita...
All pathways stored in KEGG can be downloaded in its own XML format named KGML, which is supported by VisANT, a software tool for pathway visualization (Hu et al., 2008b) and indirectly by Cytoscape using scripting plugins. In addition, metabolic pathways are available in BioPAX Level 1, which was especially designed for metabolic reactions, as well as in SBML. For converting KEGG metabolic pathways to SBML, a tool called KEGG2SBML (http://sbml.org/Software/KEGG2SBML) was developed.

KEGG data can also be accessed using the KEGG API or KEGG FTP. Moreover, to make use of the KEGG resources, several applications exist. KegArray, for example allows the analysis of microarray data in the context of KEGG pathways.

### 2.3. WikiPathways

A recently developed resource for pathway information that strongly differs from other pathway repositories is WikiPathways. WikiPathways is an open source project based, like Wikipedia, on the MediaWiki software (Pico et al., 2008). It serves as an open and collaborative platform for creation, edition and curation of biological pathways in different species.

WikiPathways aims to achieve a public commitment to pathway storage and curation by keeping pathway creation and curation processes simple. While the curation process of the previously described databases is subjected to experts, any user with an account on WikiPathways can create new pathways, and edit already existing ones.

The pathway entities are linked to reference databases, based on the criteria of the editor. Hence, the identifiers depend on the chosen reference database and can therefore differ between pathways and even within a single pathway.

Pathways in WikiPathways can be browsed by species and categories, e.g. Metabolic Process. They can also be searched using gene, protein or pathway name or any free text query. In addition, pathways can be programmatically accessed through a web service (http://www.wikipathways.org/index.php/Help:WikiPathways_Webservice).

For pathway data exchange, WikiPathways does not use standard formats like BioPAX or SBML, but offers a much simpler representation called GenMAPP Pathway Markup Language (GPML) that is compatible with visualization and analysis tools such as Cytoscape, GenMAPP (Salomonis et al., 2007) and PathVisio (van Iersel et al., 2008). The use of GPML is in agreement with the community annotation nature of the project, since it offers a simple pathway representation and several functionalities for building network diagrams. However, interoperability with other pathway databases is impeded, and substantial efforts towards combining WikiPathways with the other pathway repositories will be required.

In this regard, some approaches with the objective of conversion between GPML and standard pathway exchange formats, such as SBML and BioPAX, are under development (Evelo, 2009). In addition, KEGG pathways in KGML format are also available in GPML format ready for download (http://www.pathvisio.org/Download#Step_3) or can be converted into GPML (http://www.bigcat.unimaas.nl/tracprojects/pathvisio/wiki/KeggConverter).

The exponential growth of biological data poses a challenge to the high quality annotation and curation of databases. In this scenario, the use of wikis for community curation of biological data has emerged in the last years with the goal of increasing quality of data annotation by combining knowledge from multiple experts (Giles, 2007; Hu et al., 2008a; Waldrop, 2008). However, their success will strongly depend on the commitment of the community and WikiPathways authors claim that the initiative represents an experiment, in which the “community curation” approach is being tested (Pico et al., 2008). Thus, WikiPathways can be seen as a complementary and enhancing source of information for the major pathway databases like Reactome or KEGG.

In contrast to the aforementioned databases, the systems described below combine diverse pathway repositories, and can be seen as first attempts towards integration of pathway information from various sources.

### 2.4. Nature Pathway Interaction Database (PID)

The Pathway Interaction Database (PID) contains data on cell signalling for human (Schaefer et al.,
2009). PID combines three different sources, the NCI-curated pathways that are obtained from peer-reviewed literature, as well as pathways imported from Reactome and BioCarta. Similar to Reactome, PID structures pathways hierarchically into pathways and its sub-pathways that are called sub-networks in PID.

The PID data model is based on molecular interactions in which input biomolecules are transformed into output biomolecules. Each process can be promoted or inhibited by regulators. Biomolecules are proteins, RNA, complexes or small molecules. DNA is not part of the PID data model and only output RNA and regulator are represented in transcriptional processes. Each protein is cross-referenced to UniProt, RNA to Entrez Gene, small molecules to the Chemical Abstracts Service (CAS) registry number and complexes are annotated with GO terms. Different states of biomolecules such as “active/inactive” or “phosphorylated” are part of the annotations of the biomolecule. Cellular location, biological processes and molecular function of the entities are cross-linked to GO. Moreover, interactions are annotated with the supporting literature or other evidence, such as inferred from array experiment (Schaefer et al., 2009).

Pathways can be browsed and queried using gene or protein identifiers, such as Entrez Gene identifier, UniProt accession numbers or HUGO gene symbols, as well as biological process terms from GO, among others. The system returns available results from each of the three sources. Moreover, PID offers advanced queries. The connected molecules search option allows finding a possible path between two or more molecules. In the batch query, the user can upload lists of gene or protein identifiers and obtain a list of pathways ranked by the probability of including the entities of the query list. Using this application, pathways overrepresented in a set of genes, e.g. derived from microarray expression experiments, can be obtained.

PID provides different pathway representation formats including BioPAX Level 2 and a PID proprietary XML format. Data from Reactome is directly imported using the BioPAX Level 2 format and is regularly updated. Since not all entities or events stored in Reactome can be presented in BioPAX Level 2, some information is lost during the import. However, this might be avoided once BioPAX Level 3 will be released. The BioCarta data is manually assigned to the PID data model, as BioCarta only offers a graphical cartoon view of the pathways and does not provide computer readable download format.

2.5. Pathway Commons

Pathway Commons is a compilation of the public pathway databases Reactome, PID and Cancer Cell Map as well as protein-protein interaction databases such as HPRD (Mishra et al., 2006), HumanCyc, IntAct (Kerrien et al., 2007a) and MINT (Zanzoni et al., 2002). Herein, the pathway hierarchies of Reactome and PID are conserved.

Pathway Commons serves as an access point for a collection of public databases and provides technology for integrating pathway information. Pathway creation, extension and curation remain the duty of the source pathway databases. As a consequence, entries in Pathway Commons are cross linked to their source database, and links to external databases rely on the source database.

A regular search is provided and a filter can be set restricting the results to source and organism. Furthermore, Pathway Commons provides a web service API for an automatic access of the data. Additionally, cPath, a Java open-source software for aggregating, storing and querying pathway data, is offered. One of its key features is the identifier mapping system. It handles mapping tables of equivalent entities, such as the UniProt and the RefSeq accession number of proteins. These tables can in principle be used to integrate data from diverse sources that use different identifiers. Moreover, the system can straightforwardly be extended including self-created mapping tables. PSI-MI (Kerrien et al., 2007b) and currently BioPAX Level 2 exchange format are supported. Furthermore, the complete Pathway Commons database can be automatically accessed using the Pathway Commons plugin in Cytoscape.

The systems presented above allow the access to a wide range of data on biological pathways. However, there is overlap in the information offered by different databases. On the contrary, for specific pathways some databases offer more accurate and complete information than others. Hence, the user might have difficulties in choosing the right database and in dealing with redundancies and inconsistencies among the pathways. The integration initiatives exemplified by PID and Pathway Commons are attempts to solve these problems. However, the intended integration is not trivial since the data is fragmented and stored in databases that may differ in the representation of the biochemical reactions, as well as in the coverage and accuracy of annotations. In addition, often data is not provided in interchangeable formats hampering the
3. Case study: EGFR signalling

The epidermal growth factor receptor (EGFR) signalling cascade is one of the best studied and most important signalling pathways in mammals. It regulates cell growth, survival, proliferation and differentiation. Recently, a detailed and comprehensive map of the EGFR signalling pathway has been reported (Oda et al., 2005). Since the map was manually built by experts from literature, it can be seen as a reference representation of the pathway. Other reference maps of important signalling pathways have been reported in the last years (Calzone et al., 2008; Herrgard et al., 2008; Oda & Kitano, 2006), providing the scientific community with comprehensive maps that can be used for modelling, which in turn will shed light on important aspects of cell signalling. However, these initiatives constitute huge efforts and, as judged by the limited number of already available maps, there is a lag between the amount of data available in public databases and the availability of such references map. Hence, we argue that public pathway databases could be used to build such reference maps of signalling pathways. Most pathway databases are also developed by experts in the field and constitute repositories of high quality data, with the additional advantage of being already represented in machine readable formats that could, in principle, be easily and automatically retrieved, analyzed and fed into modelling software tools.

We selected the EGFR pathway (Oda et al., 2005), hereafter referred to as EGFR map, as a “gold standard” to evaluate the completeness and accuracy of public pathway databases in the representation of the reactions that are part of the EGFR signalling (Figure 1). We based our selection on the following reasons: (i) signalling through EGFR has been studied for more than 40 years and a lot of information on the reactions is already available (Citri & Yarden, 2006); (ii) it has been carefully curated by experts; (iii) it constitutes an excellent example of crosstalk between different signalling events, thus allowing evaluation of the coverage of crosstalks in the public databases and the ability of network analysis tools to retrieve and combine networks in a meaningful manner; (iv) the study of signalling through EGFR has important implications for understanding several cancer types and the development of new therapeutic strategies. Several computational models have been reported on different aspects of EGFR signalling (Birtwistle et al., 2007; Borisov et al., 2009; Hornberg et al., 2005; Kholodenko et al., 1999; Li et al., 2009; Schoeberl et al., 2002). However, it is worth mentioning that, to our knowledge, no model for the whole EGFR map has been reported up to now.

In the following paragraphs and in Figures 1-3 we use the same notation of entities as in the SBML file of the EGFR map (Oda et al., 2005). The EGFR map is based on more than 240 publications and contains several crosstalks between the EGFR downstream signalling and other signalling pathways. In the EGFR map, depicted in Figure 1, entities are clustered according to their cellular location and function. The functional units comprise receptor endocytosis, recycling and degradation, small GTPase signalling, MAPK cascade, PIP signalling, cell cycle, Ca2+ signalling and GPCR-mediated transactivation. Seven phenotypic outcomes of EGFR signalling are depicted: ErbB endocytosis, ErbB degradation, apoptosis, actin reorganization, cell cycle, gene transcription and mitogenesis/tumorigenesis.

In order to address the completeness and accuracy of pathway information available in public databases and its automatic retrieval, we tried to recover the complete EGFR map. For this purpose we queried Reactome version 26 with the term “EGFR” and its UniProt identifier “P00533” and downloaded and visualized the retrieved pathways. Reactome was chosen as it is currently the most detailed pathway repository, and utilizes a data model that accommodates different types of biochemical reactions. For visualization we chose Cytoscape due to its user-friendly visualization capabilities and its network analysis tools. In order to map the entities found in Reactome to those in the EGFR map, a mapping through standard identifiers was performed. We compared the original EGFR map with the EGFR pathway recovered from Reactome (in BioPAX format) and coloured entities and reactions according to their representation in both resources (see Figure 1). Red entities are found to be identical in the EGFR pathway in Reactome and the EGFR map. Purple connotes entities that could be recovered from Reactome but that are differently represented (for instance, if a single protein instead of a complex is described to take part in a reaction).

Only a small proportion of the original EGFR reactions could be recovered from Reactome, and most of them are directly related to signals coming from downstream EGFR signalling. Most of the reactions related to other signalling cascades are not connected with EGFR signalling in Reactome and could therefore not be recovered. Regarding the associated phenotypes, only two were found in Reactome: EGFR endocytosis and EGFR degradation. However, both mechanisms are described in a slightly
different manner, in some cases even with more details than in the original EGFR map. In a second step, we tried to extend the EGFR map by querying Reactome with key entities found in the EGFR map in order to complete signalling cascades such as the GPCR signalling or the MAPK cascade that are missing in the EGFR pathway in Reactome. All pathways that were added are listed in Table III. We used additional colours to depict the entities that were recovered in this extension process. Green was used for entities found in Reactome and the EGFR map and turquoise was used for entities that differed in their representation in both sources. By this extension, we were able to recover four of the five missing phenotypes: actin reorganization, apoptosis, cell cycle, and the transcription of target genes. However, some reactions were still missing in the information recovered from Reactome and in some cases, gaps or contradictions appear impeding an automatic integration (Figure 2). In this example, reactions in which ERK1 and ERK2 participate are first separately described and later the representation switches towards a combined ERK1/2 entity.

Regarding the reactions that give rise to regulatory loops in the EGFR map, only some of them could be recovered from Reactome. For instance, although the reaction that involves cleavage of pro-HB-EGF by ADAMs is described, its regulation by Pyk2 and c-Src is not included and therefore this positive feedback loop is not coloured in the EGFR map. In total, three of the six negative feedback loops were detected: inhibition of EGFR by SHP1, down regulation of EGFR, and phosphorylation of SOS1 by ERK1, which leads to SOS1 inhibition.

Although most of the crosstalks between signalling cascades in the EGFR signalling could be established by the extension process, a significant number were not found because the entities that link the different cascades are missing in Reactome. For example, the important crosstalk of the Ca$^{2+}$ and the EGFR signalling by the effect of Ca$^{2+}$ on Pyk2 activity could not be recovered, as Pyk2 is not present in Reactome. Moreover, it is worth mentioning that details in some of the reactions differ between the EGFR map and the data found in Reactome. In part, this can be explained by the fact the former is based on literature curated by 2005, and version 26 of Reactome was released during October, 2008.

The extension process was achieved by searching the database with entities representing the main signalling cascades that are known to be connected with the EGFR signalling, followed by manual identification of the reactions that connect the pathways. In principle, the process of finding the connections or crosstalks between pathways could be automated using tools available in Cytoscape or Pathway Commons (cPath). The Cytoscape merging function was evaluated for this purpose. This function compares the attributes of the nodes in order to automatically connect reactions from different pathways. However, when tested on the reactions in the EGFR map, several problems arose. Most of them appeared as a result of annotation issues. For instance, Figure 3 shows two reactions in which the two IP3 entities are differently annotated. The first IP3 entity is located in the “cytosol”, whereas this cellular location annotation is missing for the second IP3 precluding the expected merging of the two reactions. Hence, an automatic integration is impeded and manual intervention is needed. Another factor that hampers finding connections between reactions or pathways is the use of combined entities. For instance, the already mentioned ERK1 and ERK2 proteins first represented as separate entities are later described as a combined ERK1/ERK2 entity (Figure 2). This problem could be solved by considering all the annotations of the nodes to decide if two entities are equivalent or not. This would allow comparing nodes that represent states of entities, for instance post-translational modified proteins, or proteins annotated with cellular locations. In summary, reconstruction of crosstalks between signalling pathways is difficult by means of the automatic tools currently available. Manual intervention is required in order to recover all reactions involved in the pathways and their crosstalks.

The case study presented here shows that a process combining automatic retrieval and manual intervention can be used to reconstruct the EGFR map in its main features. This demonstrates that current pathway databases contain a lot of detailed information, though in some cases this information is still incomplete and manual intervention is needed to obtain a complete and correct network representation containing different signalling pathways. This is especially critical for reactions that are part of regulatory feedback loops, as these determine the dynamic behaviour of the signalling pathway. Nevertheless, information obtained for individual reactions and even for some pathways is quite complete and would be accurate enough as a starting point for model building. A proposal for a strategy for the use of pathway data from public databases for network modelling is presented in Box 1.

4. Conclusions and perspectives

In this article we have reviewed the main knowledge resources of human pathways and we have evaluated the feasibility of using this information for the reconstruction of signalling pathways in a
biologically meaningful manner. Moreover, we have presented a scenario in which data from public pathway databases are directly used for modelling (see Box 1). In this regard, we have briefly discussed the main standards for representation of biological networks, BioPAX and SBML. Furthermore, we have discussed the advantages and drawbacks of current methods for pathway retrieval and integration, using the EGFR signalling as an illustrative example.

We encourage the combination of data from different pathway databases, since they are often complementary and, in this way, a better coverage of all the reactions involved in a given pathway will be achieved. However, the integration of pathways from different databases poses a challenge, since different standard formats are used and data models vary.

Even if we choose a single database as a source of pathway information, the retrieval of a signalling pathway in conjunction with its crosstalks to other signalling cascades is a difficult task. Although in this case the data model and representation formats are not an issue, there are still annotation problems that remain to be solved to allow an effective integration of reactions and pathways. In this regard, the communication of these problems to database curators by the users will be of great help to improve the completeness and quality of annotations.

There is a strong need of tools for the automatic integration of different pathways in a biological meaningful way. The analysis presented here stresses that this is not a trivial task. Since several annotation problems and inconsistencies exist, manual intervention is needed to achieve the integration. Moreover, other factors have to be considered in order to decide if two pathways can be merged: are the pathways found in the same cell type? Or, are they found at the same developmental stage of the cell? Accurate annotation of the reactions taking place in each signalling pathway will be required in order to appropriately solve these questions.

The case study on the EGFR signalling has highlighted very important issues for the practical use of pathway databases. The information obtained for individual reactions and even for particular pathways is quite complete in most of the cases and would be accurate enough as a starting point for modelling. Although we did not perform a systematic evaluation of all the reactions found in Reactome, on the basis of the results of this case study we can conclude that public databases contain accurate and quite complete information about the main processes involved in cell signalling pathways. However, for processes for which no such level of detail on the reactions is available, the representation recovered from public databases will be less complete. For example, comparison of manually created Rb/E2F pathway with data from Reactome indicated that the latter does not cover all the reactions (Calzone et al., 2008). We foresee that in the following years, the coverage of the databases will grow as well as the quality of the annotations, which will benefit the scientific community in providing a source of representations of pathways for modelling purposes.

We would like to finish by stressing the importance of the annotation and data representation issues for an effective integration of data from public pathway databases. Researchers involved in pathway annotation and in pathway modelling should engage in collaborative projects in order to take advantage of the data already available in public databases and work together on representations that fit the needs of both communities.

Acknowledgements

This work was generated in the framework of the @neurIST and the EU-ADR projects co-financed by the European Commission through the contracts no. IST-027703 and ICT-215847, respectively. The Research Unit on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB). It is also member of the COMBIOMED network. We thank the Departament d’Innovació, Universitat i Empresa (Generalitat de Catalunya) for a grant to author ABM.
References


Figure and Box legends

Box 1: Use of data from public pathway databases for modelling purposes
Most public available pathway databases offer their data in BioPAX format, which was developed for
detailed pathway representation and as data exchange format. For storing and sharing of computational
models of biological networks, SBML has emerged as standard and is supported by most modelling
software. BioPAX and SBML, the two main standards for the representation of biological networks, have
been discussed in detail by other authors (Stromback et al., 2006; Stromback & Lambrix, 2005). In Table
II, we briefly list the most important features of the SBML and BioPAX standards.

A scenario in which pathway data is directly used for network modelling is proposed here. One or more
pathways represented in BioPAX format are automatically retrieved from different databases and
imported into a pathway visualization and analysis tool. Then, integration of the different pathways can
take place in order to obtain a comprehensive and biologically meaningful representation of the network.
In addition, annotations can be added if required or structural analysis of the network can be performed.
The resulting network, which integrates the original pathways retrieved from the databases, is exported to
SBML format, and subjected to modelling. If a quantitative approach is chosen, additional information
such as rate constants are required in order to start the modelling process.

In this process, conversion between the two formats is required in order to achieve interoperability
between pathway and model representations. Some solutions are already available. The BioModels
(http://www.ebi.ac.uk/biomodels-main/) database, which contains a variety of curated models in SBML
format, offers conversion to BioPAX format. The opposite conversion, from BioPAX to SBML, would
open the possibility of modelling the pathways stored in public databases. However, the inter-conversion
between BioPAX and SBML is not trivial as both formats where developed for different purposes.
BioPAX, for instance, does not offer the possibility to store quantitative information needed for kinetic
modelling, while SBML does not represent relationships between nodes that are not needed for modelling
and that are present in BioPAX. Examples of approaches for the conversion from BioPAX to SBML are
BiNoM (Zinovyev et al., 2008), which is available as Cytoscape plugin and SyBil, which is part of the
model environment for quantitative modelling VCell (Ruebenacker et al., 2007).

Although compatibility of different pathway and network model exchange formats is still not completely
achieved, the efforts made towards this goal represent significant contributions to pathway retrieval,
integration and subsequent modelling.

Figure 1: EGFR map

EGFR map created with CellDesigner ver.2.0 (Oda et al., 2005). This map has been coloured to show the
entities and reactions that are in common between the EGFR map and information found in Reactome.
Red colour denotes that the entities and reactions are equivalent, purple connotes that they are similar, but
differ in some description details, and white is used for entities and reactions that could not be directly
found in Reactome. In a second step, the map was extended by querying Reactome with key entities
appearing in the EGFR map, which were missing in the representation of the EGFR pathway in
Reactome. After this extension process, we coloured new equivalent entities in green and new similar
entities (the ones that are differently described in both resources) in turquoise. For comparing the EGFR
map with the pathways downloaded from Reactome, the Reactome pathways have been imported into
Cytoscape and the entities and reactions have been manually compared using the node and edge search
functions of Cytoscape. As the SBML version of the EGFR map does not contain unique identifiers for
the nodes (species), all names have been first matched to Entrez Gene identifiers, which have then been
used for comparing entities between the EGFR map and Reactome. In the extension process canonical
names as well as Entrez Gene and UniPro identifiers have been used in order to retrieve all the
information available in Reactome. In some cases, when no results were obtained in this way, the search
was additionally expanded. For example, in order to find the EGFR crosstalk with the GPCR signalling,
we additionally searched for “G protein”. GPCR signalling pathways activated by S1P1, S1P2, LPA1,
LPA2, EP2, EP3, EP4 were not found in Reactome. Instead, GPCR signalling through thrombin and glucagon
receptors that are related to EGFR signalling (Buteau et al., 2003; Prenzel et al., 1999) are present in
Reactome and were incorporated into the EGFR map in order to complete the missing crosstalk.

Figure 2: Comparison of ERK signalling as found in the EGFR map and in Reactome
In Reactome, ERK1 or ERK2 are phosphorylated by MKK1 or MKK2, respectively. The same reaction is found in the *EGFR map*, but here ERK1 and ERK2 are represented in a single entity, namely ERK1/2, which combines both proteins and can be phosphorylated by MKK1 and MKK2. In Reactome, dimerization and translocation to the nucleus are still separately described for each entity. Then, the representation switches from separate entities to one combined entity. In addition, the dissociation of MKK1 or MKK2, which is needed before the dimerization of ERK can take place, is not described in Reactome. Manual intervention is needed in order to correctly map both representations.

**Figure 3: Example – annotation issue**

The two reactions “Active PLCγ hydrolyses PI4,5-P₂” (REACT_12078.2) and “IP3 binds with the IP3 receptor, opening the Ca²⁺ channel” (REACT_12008.1) are connected through the entity IP3 as the hydrolysis of PI4,5-P₂ results in DAG and IP3, and then IP3 binds to its receptor enabling the Ca²⁺ release. However, the process in which IP3, located in the cytosol, translocates to the plasma membrane to bind to its receptor is not precisely described in Reactome, leading to differences in the annotation of both IP3 entities. This precludes the automated merging of both chains of reactions and manual intervention was needed to connect the reactions correctly.
Supplementary information

SBML and PDF files of the comprehensive map of EGFR signalling with entities coloured according to their appearance in Reactome.
- EGFR_map_coloured.xml
- EGFR_map_coloured.pdf
**EGFR map**

- **phosphorylation of ERK1/2**
- **dimerization**
- **translocation to the nucleus**
- **activation of ELK1**

**Reactome**

- **missing dissociation of MKK1, MKK2 from ERK1, ERK2**
- **missing link between ERK1, ERK2 and ERK1/2**
Active PLC$_Y$ hydrolyses PI$_{4,5}$-P$_2$.

IP$_3$ binds with the IP$_3$ receptor, opening the Ca$^{2+}$ channel.
<table>
<thead>
<tr>
<th>Pathway/PPI database</th>
<th>Web-link</th>
<th>Standard exchange formats for download</th>
<th>Web service API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reactome</td>
<td><a href="http://www.reactome.org">http://www.reactome.org</a></td>
<td>BioPAX Level 2, BioPAX Level 3 (only some reactions), SBML Level 2</td>
<td>SOAP web service API, Detailed user manual available, example client in Java</td>
</tr>
<tr>
<td>KEGG</td>
<td><a href="http://www.genome.jp/kegg/pathway.html">http://www.genome.jp/kegg/pathway.html</a></td>
<td>KGML (default format), BioPAX Level 1 (only metabolic reactions), SBML (using converter), GPML (using converter)</td>
<td>SOAP web service API, Example client in Java, Ruby, Perl, Direct import into Cytoscape</td>
</tr>
<tr>
<td>WikiPathways</td>
<td><a href="http://www.wikipathways.org">http://www.wikipathways.org</a></td>
<td>GPML (default format), Converters to standards such as SBML and BioPAX are in progress.</td>
<td>SOAP web service API, Example clients in Java, Perl, Python, R</td>
</tr>
<tr>
<td>NCI / Nature Pathway Interaction Database (PID)</td>
<td><a href="http://pid.nci.nih.gov">http://pid.nci.nih.gov</a></td>
<td>PID XML (default format), BioPAX Level 2</td>
<td>Access through Pathway Commons</td>
</tr>
<tr>
<td>Pathway Commons</td>
<td><a href="http://www.pathwaycommons.org">http://www.pathwaycommons.org</a></td>
<td>BioPAX Level 2 (default format for pathways), PSI-MI (default format for protein-protein interactions)</td>
<td>HTTP URL based XML web service via cPath, Direct import into Cytoscape</td>
</tr>
<tr>
<td>Cancer Cell Map</td>
<td><a href="http://cancer.cellmap.org">http://cancer.cellmap.org</a></td>
<td>BioPAX Level 2</td>
<td>HTTP URL based XML web service via cPath</td>
</tr>
<tr>
<td>HumanCyc</td>
<td><a href="http://humancyc.org">http://humancyc.org</a></td>
<td>BioPAX Level 2</td>
<td>Access through Pathway Commons and Pathway Tools</td>
</tr>
</tbody>
</table>
### References

<table>
<thead>
<tr>
<th><strong>Comparison SBML-BioPAX</strong></th>
<th><strong>SBML</strong></th>
<th><strong>BioPAX</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Representation format</strong></td>
<td>XML (Extensible Markup Language)</td>
<td>OWL (Web Ontology Language) XML</td>
</tr>
<tr>
<td><strong>Main purpose</strong></td>
<td>Representation of computational models of biological networks</td>
<td>Pathway description with all details on reactions, components, information on cellular location etc.</td>
</tr>
<tr>
<td><strong>Entities and reactions</strong></td>
<td>Based on species and reactions (Hucka et al., 2003): - species (proteins, small molecules etc.) - reactions (how species interact) - compartment (where interactions take place)</td>
<td>Basic ontology based on three classes (<a href="http://www.biopax.org/">http://www.biopax.org/</a>): - pathway (set of interactions) - physicalEntity with subclasses such as RNA, DNA, protein, complex, small molecules - interaction with subclasses such as conversion having biochemicalReaction as subclass, etc.</td>
</tr>
<tr>
<td><strong>Number of pathways represented</strong></td>
<td>One model per SBML file</td>
<td>Several pathways per BioPAX file possible (each object has its own RDF id and is hence uniquely identifiable)</td>
</tr>
<tr>
<td><strong>Reaction kinetics</strong></td>
<td>Allows representation of kinetics including parameters for reaction rates, initial concentrations etc.</td>
<td>No kinetics as BioPAX is not meant for modelling but pathway representation</td>
</tr>
<tr>
<td><strong>Levels</strong></td>
<td>Built in levels with different versions. Each level adds new features such as the incorporation of controlled vocabularies. At the time of writing, the most stable version is SBML Level 2.</td>
<td>BioPAX Level 1: representation of chemical reactions involved in metabolism BioPAX Level 2: adds molecular interactions and protein post-translational modifications BioPAX Level 3: any kind of biological reaction, including regulation of gene expression (BioPAX L3 is at the time of writing still in release process) The BioPAX project roadmap envisages two additional levels capturing interactions at the cellular level. (<a href="http://www.biopax.org/Docs/BioPAX_Roadmap.html">http://www.biopax.org/Docs/BioPAX_Roadmap.html</a>)</td>
</tr>
<tr>
<td><strong>Pathway database support</strong></td>
<td>- Reactome - KEGG</td>
<td>- Reactome - KEGG (only BioPAX Level 1)</td>
</tr>
<tr>
<td>Category</td>
<td>Support</td>
<td>Notes</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Model database support</td>
<td>BioModels</td>
<td>BioModels (conversion from SBML to BioPAX possible)</td>
</tr>
<tr>
<td>Library for reading/writing</td>
<td>libSBML (Bornstein et al., 2008)</td>
<td>Paxtools (<a href="http://www.biopax.org/paxtools/">http://www.biopax.org/paxtools/</a>)</td>
</tr>
<tr>
<td>Software support</td>
<td>- Standard modelling software</td>
<td>Network visualization software such as Cytoscape, VisANT</td>
</tr>
<tr>
<td></td>
<td>such as CellDesigner or Copasi</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Hoops et al., 2006)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Network visualization software</td>
<td></td>
</tr>
<tr>
<td></td>
<td>such as Cytoscape</td>
<td></td>
</tr>
</tbody>
</table>
Table III: Pathways downloaded for extending the *EGFR map*

<table>
<thead>
<tr>
<th>Extension to</th>
<th>Reactome pathways downloaded</th>
<th>Reactome identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP kinase cascade</td>
<td>- RAF activation</td>
<td>REACT 2077.4</td>
</tr>
<tr>
<td></td>
<td>- MAP kinase cascade</td>
<td>REACT 634.4</td>
</tr>
<tr>
<td></td>
<td>- ERK1/2 activates ELK1</td>
<td>REACT 12406.1</td>
</tr>
<tr>
<td></td>
<td>- ERK1/2/5 activate RSK1/2/3</td>
<td>REACT 12487.1</td>
</tr>
<tr>
<td></td>
<td>- RSK1/2/3 phosphorylates CREB at Serine 133</td>
<td>REACT 12622.1</td>
</tr>
<tr>
<td>Ca²⁺ signaling</td>
<td>- Active PLCG1 hydrolyses PIP2</td>
<td>REACT 12078.2</td>
</tr>
<tr>
<td></td>
<td>- DAG stimulates protein kinase C-delta</td>
<td>REACT 12062.1</td>
</tr>
<tr>
<td></td>
<td>- IP3 binds with the IP3 receptor, opening the Ca²⁺ channel</td>
<td>REACT 12008.1</td>
</tr>
<tr>
<td></td>
<td>- Release of calcium from intracellular stores by IP3 receptor activation</td>
<td>REACT 12074.1</td>
</tr>
<tr>
<td></td>
<td>- Calcium binds calmodulin</td>
<td>REACT 12602.1</td>
</tr>
<tr>
<td>GPCR signaling</td>
<td>- Thrombin-activated activation cascade</td>
<td>REACT 57.1</td>
</tr>
<tr>
<td></td>
<td>- Glucagon signaling in metabolic regulation</td>
<td>REACT 1665.2</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>- p53-Dependent G1/S DNA damage checkpoint</td>
<td>REACT 85.1</td>
</tr>
<tr>
<td></td>
<td>- NRAGE signals death through JNK</td>
<td>REACT 13638.1</td>
</tr>
<tr>
<td></td>
<td>- Activation of BAD and translocation to mitochondria</td>
<td>REACT 549.2</td>
</tr>
<tr>
<td></td>
<td>- BH3-only proteins associate with and inactivate anti-apoptotic BCL-2 members</td>
<td>REACT 330.1</td>
</tr>
<tr>
<td></td>
<td>- Cyclin D associated events in G1</td>
<td>REACT 821.2</td>
</tr>
<tr>
<td></td>
<td>- Cyclin E associated events during G1/S transition</td>
<td>REACT 1574.2</td>
</tr>
</tbody>
</table>