Minireview

Pathway information for systems biology

Michael P. Cary, Gary D. Bader, Chris Sander

Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, Box 460, New York, NY 10021, USA

Accepted 1 February 2005

Available online 9 February 2005

Edited by Robert Russell and Giulio Superti-Furga

Abstract Pathway information is vital for successful quantitative modeling of biological systems. The almost 170 online pathway databases vary widely in coverage and representation of biological processes, making their use extremely difficult. Future pathway information systems for querying, visualization and analysis must support standard exchange formats to successfully integrate data on a large scale. Such integrated systems will greatly facilitate the constructive cycle of computational model building and experimental verification that lies at the heart of systems biology.

Keywords: Pathway data integration; Pathway database; Standard exchange format; Ontology; Information system

1. Introduction

To understand biological processes, we must integrate new observations with existing knowledge to create testable models that can be iteratively refined. This will only be successful if the vast amounts of data gathered by large-scale profiling of biological features, such as mRNA transcripts and proteins, can be efficiently integrated with data from the literature and databases for visualization and analysis.

One major source for computable data about biological processes are databases that capture information on the functional interactions of molecular species [1]. These "pathway" databases facilitate a variety of analysis and simulation techniques that can enrich our understanding of cellular systems.

While recent dramatic growth in the number of pathway databases is a great boon to biologists, it also presents several important challenges. Almost 170 "pathway" databases exist, which differ widely in form and content. This multiplicity of information sources can be daunting to researchers who simply wish to find information about genes or pathways of interest. The lack of uniform data models and data access methods makes pathway data integration extremely difficult, both mechanistically and semantically.

To address these issues, it is useful to review the current landscape of pathway data and techniques for data integration, and then to extrapolate the shape of desirable pathway information systems which flexibly and efficiently facilitate the analysis and modeling of biological systems.

2. Surveying the pathway data landscape

One abstraction that biologists have found extremely useful in their efforts to describe and understand the inner workings of cellular biology is the notion of a biomolecular network, often called a pathway. A pathway is a set of interactions, or functional relationships, between the physical and/or genetic [2] components of the cell which operate in concert to carry out a biological process. Despite tremendous variety in the cellular processes described as pathways, several pathway representation patterns are prevalent in current practice. In the Pathway Resource List, a catalog of almost 170 pathway databases (see http://cbio.mskcc.org/prl), we use these patterns to group pathway databases into four major, slightly overlapping categories: metabolic, signaling, protein interaction, and gene regulation. A description of the major features of these categories provides an overview of the current pathway data landscape.

Metabolic pathway databases generally contain detailed data models that represent a pathway as a series of biochemical reactions, focusing mainly on the chemical modifications made to the small molecule substrates of enzymes (Fig. 1A). Many metabolic pathways have been mapped to the molecular level of detail since the 1950s or earlier and metabolic pathway databases are the earliest and perhaps the best-known. Metabolic databases generally do not represent higher order cellular processes, such as gene regulation.

Metabolic databases predominantly contain prokaryotic pathways, about which rich datasets have been collected. A few metabolic pathway databases, for example KEGG [3], the BioCyc database family [4] and others [5], map pathways from well-studied organisms onto other organisms via functional annotations, such as Enzyme Commission numbers [6], and orthology relationships, but these approaches are imperfect and the resulting pathways often contain a number of gaps, i.e., missing steps in a chain of biochemical reactions. Gap-filling algorithms attempt to address this problem [7].

Signaling pathways propagate information from one part or sub-process of the cell to another, often via a series of protein covalent modifications, such as protein phosphorylation. Dysregulation of biological processes by aberrant signaling pathways causes many common diseases, such as cancer and

E-mail address: pathways_febs@cbio.mskcc.org.

0014-5793/$30.00
doi:10.1016/j.febslet.2005.02.005

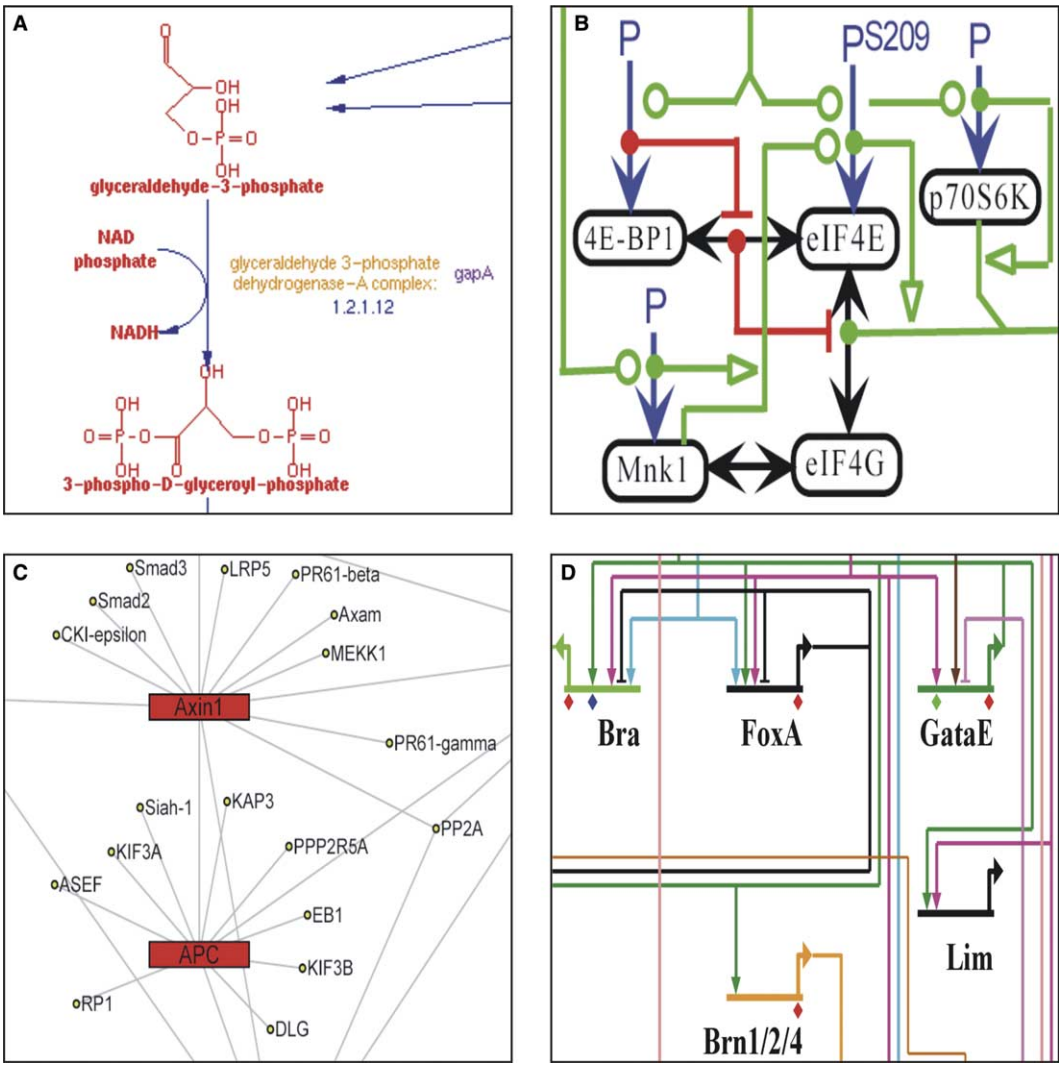

Fig. 1. Common alternative representations of pathway data. (A) Section of the glycolysis 1 pathway diagram from EcoCyc [50], drawn in high detail mode, showing a single biochemical reaction. Blue arrows depict biochemical conversion of substrates to products. The conversion arrows are labeled with the catalyzing enzyme using gold text. (B) Section of a molecular interaction map from the eMIM resource [51] showing regulation of hypoxia-responsive genes. Diagram shows phosphorylation events (blue arrows originating in blue letter P's; phosphorylation sites, if known, are abbreviated in superscript, e.g., S209 = serine 209), inhibitory relationships (red flat-headed arrows), enzymatic stimulation of events (green lines ending in open circles), binding interactions (black double-headed arrows), and non-specific stimulation of events (green arrows). Proteins are shown in black ovals, nodes (filled circles) placed on lines represent the products of processes; e.g., the node on the binding interaction arrow between eIF4E and eIF4G represents the eIF4E:eIF4G complex. (C) Section of the WNT pathway diagram from HPRD [21]. Proteins identified as important components of the pathway are shown as red boxes, other proteins are depicted as small yellow circles. Protein–protein interactions are drawn as edges between proteins. (D) Section of the endomesoderm gene network in the BioTapestry network viewer (see http://www.biotapestry.org). Genes are shown as short, thick horizontal lines. Gene products are represented as short vertical arrows originating at genes and ending in right angles. Activating and inhibitory relationships are shown as normal and flat-headed arrows, respectively, drawn from gene products to regulated genes.

diabetes [8,9]. Though not as well-established as metabolic pathway databases, signaling pathway databases are being actively constructed by a number of groups.

As many signaling pathways are present only in multi-cellular organisms, signaling databases tend to focus on eukaryotes. These organisms are much more complex and less well-studied than some bacteria and their signaling pathways appear to be more diverse than metabolic pathways. Accordingly, signaling pathway databases tend to use higher level abstractions compared to metabolic databases (Fig. 1B). For example, CSNDB [10], TRANSPATH [11] and others [12,13], often forego detailed description of the biochemical reactions involved in signaling and instead use generic concepts of activation and inhibition.

Protein interaction databases contain by far the largest number of interactions of any type of pathway database. Large amounts of protein interactions (protein–protein, protein–DNA, etc.) are generated by various large-scale experimental methods, unlike metabolic and signaling pathway data, which are generated primarily by traditional small-scale experimental techniques [14]. A well-known problem with most high throughput methods of detecting molecular interactions is the high rate of false positive results they generate [15]. Protein interactions detected by these methods should therefore be treated with less confidence until they have been verified by repeated observations or orthogonal experiments [16], and storing experimental evidence for each interaction is important for most protein interaction databases.