Opinion TRENDS in Biochemical Sciences Vol.32 No.7

Comparing classical pathways and modern networks: towards the development of an edge ontology

Long J. Lu[1,6], Andrea Sboner[1], Yuanpeng J. Huang[4], Hao Xin Lu[1], Tara A. Gianoulis[1], Kevin Y. Yip[2], Philip M. Kim[1], Gaetano T. Montelione[4,5] and Mark B. Gerstein[1,2,3]

[1] Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA
[2] Department of Computer Science, Yale University, 51 Prospect Street, New Haven, CT 06511, USA
[3] Program in Computational Biology and Bioinformatics, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA
[4] Department of Molecular Biology and Biochemistry, Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, NJ 08854, USA
[5] Department of Biochemistry, Robert Wood Johnson Medical School, UMDNJ, Piscataway, NJ 08854, USA
[6] Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, 3333 Burnet Avenue, Cincinnati, OH 45229, USA

Pathways are integral to systems biology. Their classical representation has proven useful but is inconsistent in the meaning assigned to each arrow (or edge) and inadvertently implies the isolation of one pathway from another. Conversely, modern high-throughput (HTP) experiments offer standardized networks that facilitate topological calculations. Combining these perspectives, classical pathways can be embedded within large-scale networks and thus demonstrate the crosstalk between them. As more diverse types of HTP data become available, both perspectives can be effectively merged, embedding pathways simultaneously in multiple networks. However, the original problem still remains – the current edge representation is inadequate to accurately convey all the information in pathways. Therefore, we suggest that a standardized and well-defined edge ontology is necessary and propose a prototype as a starting point for reaching this goal.

Uniting classical pathways and modern networks

In biology, a pathway refers to a sequence of reactions, usually controlled and catalyzed by enzymes, by which one organic substance is converted to another. Biological pathways are an important component of systems biology. The classical representation of these pathways provides varied, mechanistic associations between many proteins. Conversely, modern high-throughput (HTP) experiments and large-scale databases have given rise to standardized networks that provide a somewhat different perspective on pathways. By combining and comparing these perspectives, classical biochemical pathways can be embedded into large-scale networks. This reveals two problematic issues with classical pathways: (i) the components included and their exact symbolic representation (e.g. the meaning of each arrow) in the same pathway that has been documented in different databases are often inconsistent; and (ii) pathways are isolated from one another in classical representations, which de-emphasizes crosstalk. By contrast, embedded pathways offer completely uniform representations and relate network statistics such as average degree or diameter consistently. However, they are more limited in the level of detail of the mechanistic biochemistry that they can convey. As more diverse types of HTP data become available, it will be possible to embed classical pathways simultaneously in many large-scale networks, effectively merging both approaches. To accomplish this, a precise edge (or arrow) ontology needs to be defined. For illustrative purposes, we propose a prototype of ontology that provides an unambiguous representation of the edges connecting biomolecules and that also describes higher-level relationships among edges. Here, we demonstrate the usefulness of the simple-edge ontology on four diverse types of pathways. We do not intend to provide a complete ontology here but, rather, we want to stimulate people working in this field to continue building upon existing knowledge until a complete ontology is achieved.

Pathway databases and limitations

During the past decade, an increasing number of pathway databases have been established to document the ever-expanding knowledge regarding established pathways. Some of these pathway databases are organism specific. For example, EcoCyc [1] describes the genome and the biochemical machinery of *Escherichia coli* (K12 MG1655). A few other pathway databases focus on a specific type of disease or disorder, for example, The Cancer Cell Map (http://cancer.cellmap.org) or GOLD.db [2]. The majority of these pathway databases cover a certain functional area that occurs in multiple organisms. Furthermore, such databases can often be approximately divided into three categories: (i) those containing metabolic pathways (e.g. KEGG [3], WIT [4], BioCyc [5], MetaCyc [6] and GenMAPP [7]); (ii) those containing signal-transduction (signaling)

Corresponding author: Gerstein, M.B. (Mark.Gerstein@yale.edu).
Available online 20 June 2007.

 0968-0004/$ – see front matter doi:10.1016/j.tibs.2007.06.003

pathways {e.g. BioCarta (http://biocarta.com), STKE (http://stke.sciencemag.org), Pathways Knowledge Base (http://ingenuity.com) and Reactome [8]}; and (iii) those containing both (e.g. KEGG, BioCarta and Reactome). Excellent recent reviews on these pathway databases can be found elsewhere [9,10].

Although the afore-mentioned databases provide valuable resources for studying associations between proteins, they are hampered by several limitations. First, the same pathways documented in different databases are often inconsistent. In many cases, a pathway is described by including a few core components first. The decision of whether to include additional components in the given pathway is usually empirically determined, based on the expert curators' knowledge and experience. Therefore, the boundary of a pathway is usually vague. The consequence is that the number of components in the same pathway in different databases varies greatly (See Supplementary Table S1).

Second, these pathways are isolated in classical representations. This is the consequence of the traditional reductionist approaches to molecular biology, whereby genes and pathways are investigated as isolated entities. However, from the perspective of modern systems biology, the interactions between biological pathways must be studied to understand how biological systems function. On the systems level, the crosstalk between pathways seems to be particularly important but lacks substantial study. Although there have been efforts to integrate them, such as in the Boehringer Mannheim Biochemical Pathways wall chart, many aspects of the relationships between pathways have yet to be systematically identified and incorporated.

Third, the classical representations of pathways use symbols that lack a precise definition. The same symbol is often used to represent a variety of functions. For example, arrows are used to represent direct interactions in some circumstances but, in others, they are also used to represent translocation to a different subcellular compartment. Although this might not cause problems for laboratories that focus on individual pathways, these notations must be precisely defined to perform analyses on pathways on larger scales. A structured vocabulary or ontology of these symbols should be developed to ameliorate this problem.

Recent advent of network biology

A particularly novel concept in the post-genomic era is the idea that a living cell can be viewed as a complex network of biomolecules. Indeed, a biomolecular network can now be rendered as a collection of nodes and edges. Nodes represent biomolecules such as proteins, genes and metabolites, whereas edges represent the types of associations between two nodes, such as physical interactions and co-expression of mRNAs. The combined functions and interactions between these networks constitute the behavior of the cell. Mapping and understanding biomolecular networks represents the first step towards modeling how a cell actually operates.

As a result of recent genome-wide HTP experiments, including large-scale yeast two-hybrid screens and microarray experiments, many types of networks have been mapped, including protein–protein interaction, expression, regulatory, metabolic and signaling networks. For example, protein–protein interaction networks have been experimentally determined in *Saccharomyces cerevisiae* [11–15], *Caenorhabditis elegans* [16], *Drosophila melanogaster* [17], *Homo sapiens* [18,19], *Plasmodium falciparum* [20] and *Helicobacter pylori* [21]. The availability of such well-mapped networks has enabled us to compare and contrast them in terms of global and local topology, in addition to relating the structural properties of these networks to protein properties such as function and essentiality.

Topological analysis of networks provides quantitative insight into their basic organization. Different network statistics have been designed to capture the characteristics of network topology (see Supplementary Table S2). Despite the seemingly vast differences between biomolecular networks, they are found to share common features with respect to network topology. Barabási *et al.* [22] proposed a 'scale-free' model in which the degree distribution in many large networks follows a power-law distribution [$P(k) \approx k^{-\gamma}$]. What is remarkable about this distribution is that, whereas most of the nodes within these networks have few links, a few of these nodes, classified as hubs, are exceptionally well-connected. Concurrently, Watts and Strogatz [23] found that many networks also have a 'small-world' property, meaning they are defined as being both highly clustered and containing small characteristic path lengths.

Network analysis has provided new quantitative insights into protein properties, cellular dynamics and other biological problems. For example, research has shown that hubs in a network are more likely to be essential proteins, and there is debate over whether hubs tend to evolve slower [24–26]. Furthermore, different motifs have been implicated in different stages of dynamic transitions of a network [27].

Comparisons between classical and embedded pathways

Large-scale networks can be constructed using different types of data from HTP experiments: protein–protein interaction networks from yeast two-hybrid screens, and co-expression networks from microarray experiments provide apt examples of this. For each classical pathway, the corresponding sub-network can be extracted from the entire network by mapping the core components in the classical pathway onto the network of biomolecules. From a network point of view, this mapping can also be regarded as embedding pathway components into the network. To differentiate from the classical pathways, we refer to these sub-networks as embedded pathways (Figure 1). The core components of a classical or embedded pathway are defined as the biomolecules in the KEGG pathway diagram. For this review, we used KEGG because of its high quality, as pointed out by Wittig *et al.* [28].

The Notch pathway can be used as an example to illustrate embedded pathways because of its elegance and simplicity. The Notch signaling pathway is a highly conserved pathway for cell–cell communication that is involved

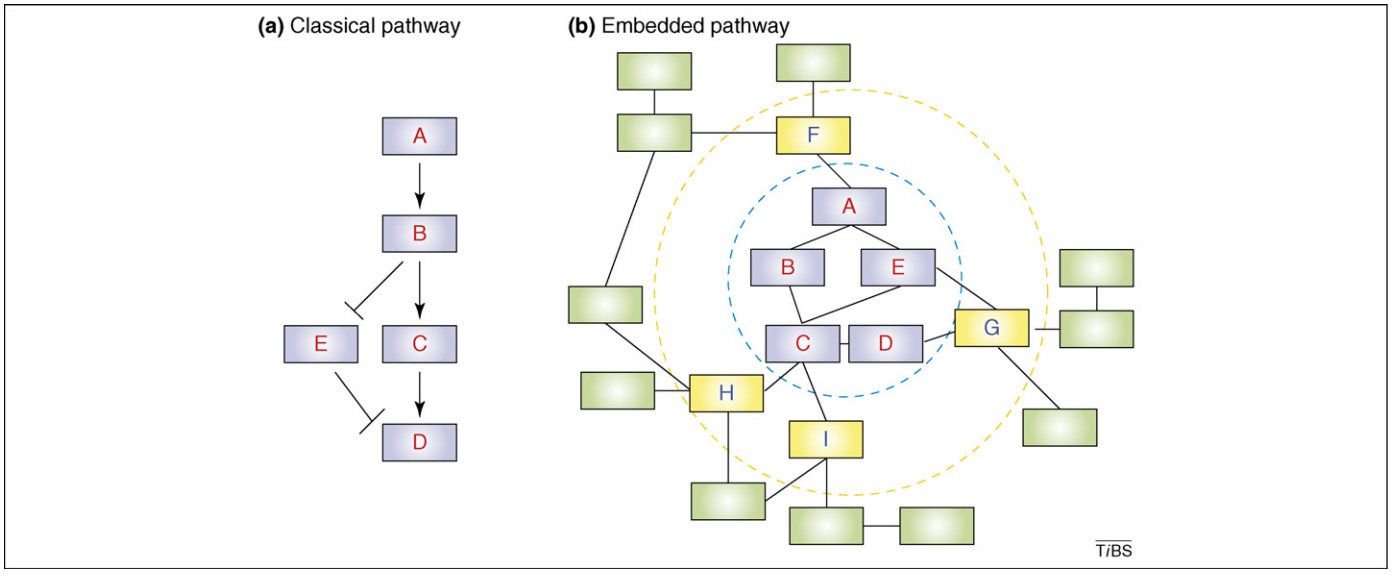


Figure 1. Classical versus embedded pathway. (**a**) Because of their wide use in textbooks, biochemists are probably most familiar with the classical representation of pathway, in which reactions and interactions are presented in a typically linear manner from the input to the output. In classical pathways, the edges are added according to expert curators' knowledge. (**b**) The recent phenomenon of showing pathways in a systems biology manner results in an embedded pathway. In this instance, the interaction between components is not necessarily linear, and components potentially involved with but outside the immediate pathway can also be shown. In this type of representation, the edges are mapped according to data from HTP experiments such as yeast two-hybrid screens of protein–protein interactions or from large-scale databases. In the example shown here, a core embedded pathway contains the same set of core components (A–E, blue nodes) as in the classical pathway in part (a), including the edges linking them together. However, the extended embedded pathway (b) also contains the nodes (yellow) that are immediately linked to the core components. Note that different numbers of edges have been intentionally drawn among the blue nodes to emphasize the potential for the occurrence of differences between classical and embedded pathways. In this example, A is shown to interact with E in the embedded pathway, whereas no such interaction is shown in the classical pathway. This is because A does not inhibit or cause E to perform a chemical reaction and, therefore, no representation of this is required in the classical pathway. However, the interaction in the embedded pathway could indicate that A functions as a scaffold for E within the pathway but does not necessarily imply that it causes E to perform a reaction.

in the regulation of cellular differentiation and proliferation. We constructed core and extended embedded pathways by collecting the 22 core-protein components listed in KEGG and mapping them onto the large-scale protein–protein interaction network deposited in the Human Protein Reference Database (HPRD) [29] (Figure 1). The HPRD interactions are manually curated by expert biologists to reduce errors.

Comparisons between classical and core embedded Notch pathways reveal several differences. First, the classical pathway contains directed and undirected edges (Figure 2a). Directed edges often represent activations, such as the edge between Delta and Notch. They also represent translocation to a different cellular compartment, for example, the edge between Notch and the Notch intracellular domain (NICD). Undirected edges often represent an interaction between two components, such as the edge between CSL (recombination signal-binding protein for immunoglobulin κJ region-like) and SKIP (SNW domain-containing 1). By contrast, the edges between components in the embedded pathway are uniform (Figure 2b). In this case, they are protein–protein interactions. Although the edge representation in the core embedded pathway is more consistent, it loses information encoded in classical pathways.

Second, although most of the edges are common between both representations, some appear only in one representation. The core embedded pathway also reveals 12 new interactions that are not found in the KEGG classical pathway. Conversely, two edges in the KEGG pathway are not present in the core embedded pathway: between Notch and Dishevelled (DVL) and between Notch and TACE (ADAM metallopeptidase domain 17), which indicates either that the protein–protein interaction map is incomplete or that these interactions take place through an intermediate (Figure 2a and b).

Compared with the classical pathway, the core embedded pathway has two advantages. First, it can indicate which isoform is responsible for an interaction. For example, in the interaction between Notch and Numb, the embedded pathway identifies that Notch1 (Entrez ID: 4851) – but not the other three isoforms – interacts with Numb (Figure 2c). By contrast, the current version of the classical pathway collapses multiple protein isoforms into one single node.

Second, extended embedded pathways can systematically represent new components involved in classical pathways. The extended embedded Notch pathway identifies 218 new proteins that are potentially involved in the Notch pathway by extracting the immediate interacting partners of these core components (Figure 2d). It is increasingly evident that the Notch pathway is subject to a wide array of regulatory influences, from those that affect ligand–receptor interactions to those that govern the choice of Notch target genes [30,31]. For example, the classical Notch pathway in KEGG shows that DVL inhibits Notch. In the HTP networks, Notch and DVL do not interact directly but through an intermediate protein, namely, glycogen synthase kinase 3β (GSK-3β). DVL and GSK-3β are known to be involved in the Wnt pathway. The interaction of Wnt with Frizzled receptors activates a cascade for which DVL is required. Activated DVL inhibits GSK-3β [32]. The relationship between GSK-3β and Notch has been found by Espinosa *et al.* [33], who report that GSK-3β phosphorylates Notch2 both *in vitro* and *in vivo*. They suggest that GSK-3β can partially mediate crosstalk between Wnt and Notch pathways.

Despite these advantages, the embedded pathway suffers substantial information loss by restricting the edges to describing physical interaction. One way to circumvent this problem would be to overlay additional types of large-scale data onto the network by defining different types of edges. For example, it has been found that Notch down-regulates Presenilin 1 (PSEN). This interaction is particularly interesting because PSEN is a component of the γ-secretase complex, which cleaves the intracellular domain of Notch, triggering the rest of the pathway. By laying the regulatory network on top of the protein–protein interactions, this feedback loop is highlighted [34].

Relating network properties in embedded pathways

Because of the heterogeneity of the edges and the incomplete nature of classical pathways, it is difficult to relate the mathematical quantities of modern network biology to these pathways. However, the same task becomes straightforward when applied to the embedded pathways created by mapping the core components of classical pathways onto large-scale networks. We provide an illustrative example as Supplementary material, showing how the topological quantities in modern network biology can lead to new insights into biochemical pathways. We found that signaling pathways from metabolic pathways have significantly different network topologies (Supplementary

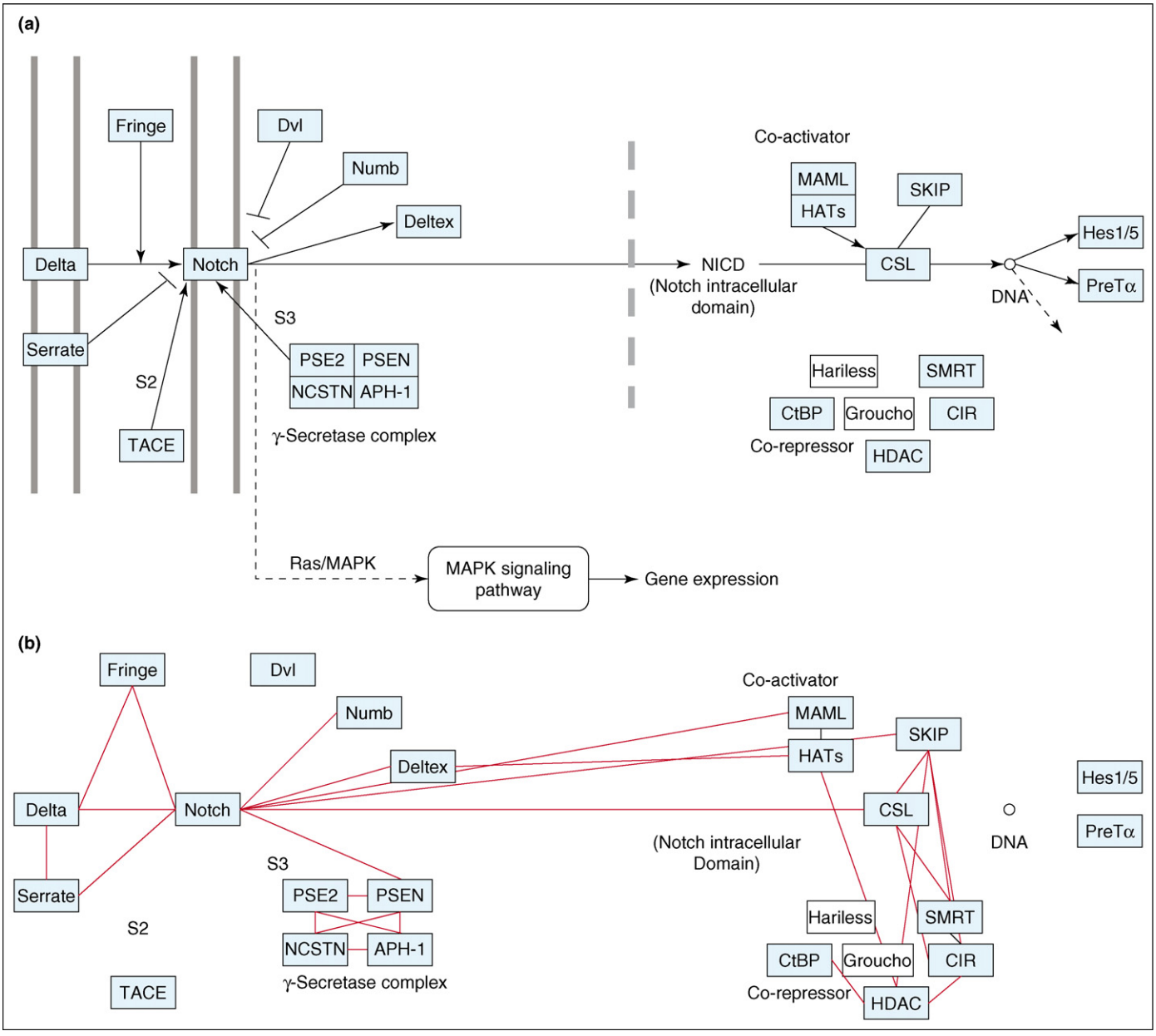


Figure 2. Comparisons of classical and embedded representations of the Notch signaling pathway. By highlighting the differences between the two types of representations, this figure demonstrates how embedding classical pathways into large-scale networks might generate new insights. In all cases, white nodes refer to reference pathway elements that are not present in the HPRD. Blue nodes are core components in HPRD and yellow nodes are the extended components mapped from HPRD. (**a**) The Notch signaling pathway as illustrated in KEGG. Both directed and undirected edges are used, and exactly the same type of edge is often assigned multiple meanings, for example, directed edges (i.e. arrows) represent activations (e.g. between Delta and Notch) and they also represent translocation to a different cellular compartment (e.g. between Notch and NICD). (**b**) The Notch pathway mapped onto interaction networks. The red edges between components are uniform, all representing protein–protein interactions between the core (blue nodes) components.

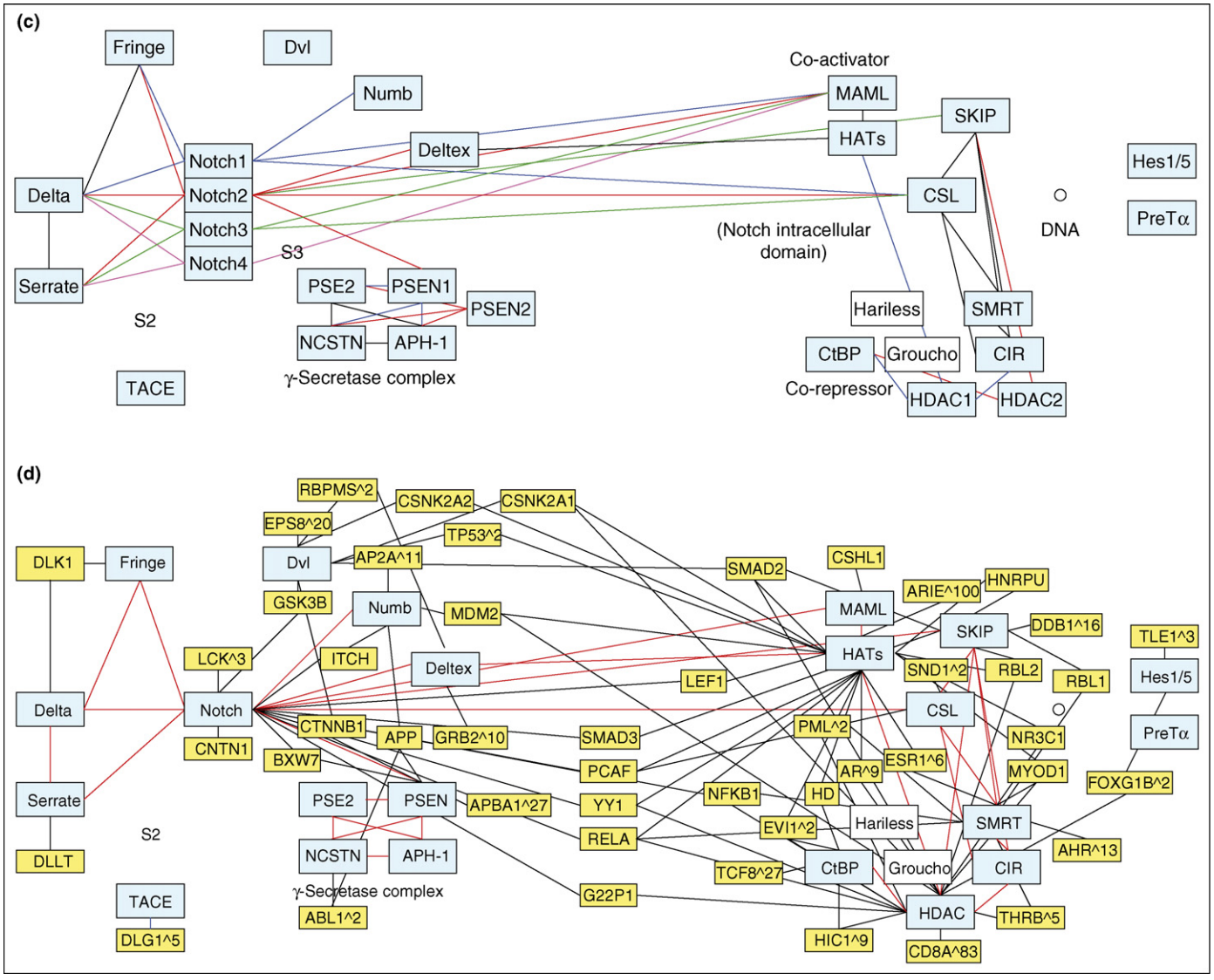
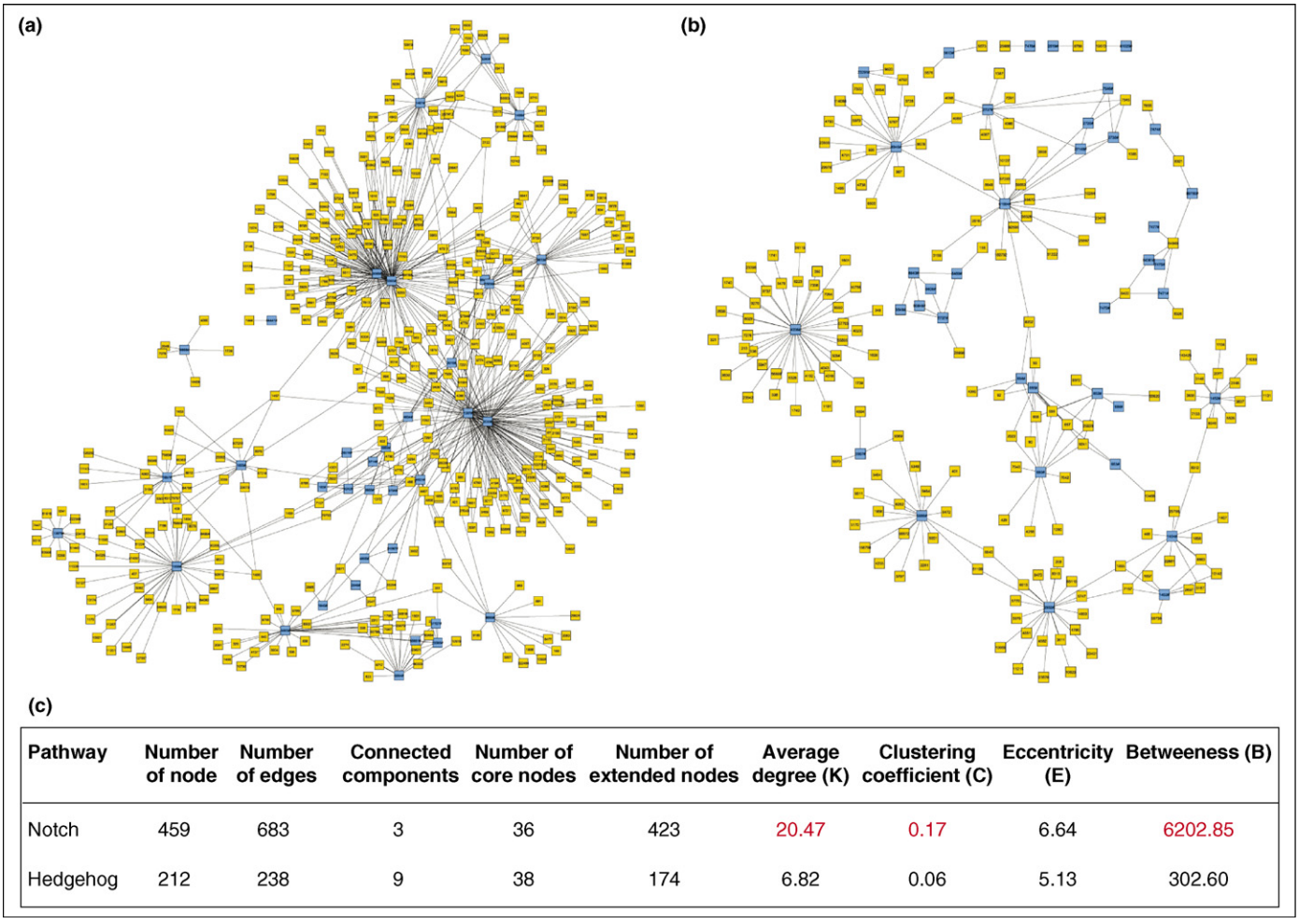


Fig. 2. Cont. (**c**) Multiple nodes in the embedded Notch pathway. Different colors are used to differentiate the interactions between isoforms. Blue edges are used to represent the interactions between Notch1 and its interacting partners, red edges represent the interactions between Notch2 and its interacting partners, and so on. (**d**) First neighbors of Notch pathway proteins. Blue nodes represent the core components of the Notch pathway, and yellow nodes represent their interacting proteins. Red edges connect interactions between core components, and black edges connect interactions between core and extended components.

Table S3 and Figure S1). This difference has enabled us to successfully differentiate signaling pathways from metabolic pathways.

It is also interesting to note that the topological quantities between two of the same type of pathway (signaling or metabolic) can be different even when they contain a similar number of core components, as illustrated by Notch versus Hedgehog extended embedded pathways (Figure 3). Do these differences reveal anything about the underlying mechanisms of Notch and Hedgehog? Initially, we must consider the possibility that these differences are merely artifacts, that is, that the protein-interaction network is incomplete and, as the map expands, such differences will disappear. Although this is plausible, an alternative explanation is that these differences reflect real biological differences. This might be explained by the different regulatory mechanisms used by the two pathways. For example, Notch pathway might be subject to more regulatory influences. Indeed, although our analysis indicates that whereas the core components of Notch embedded pathway have an average of ~21 interacting partners, those of the Hedgehog embedded pathway have just ~7 (Figure 3).

In addition to global inferences, topological measures can be used to identify nodes of particular importance or function (see Supplementary Table S6 for detailed definition of network topologies). For instance, hubs (nodes of high degree) in regulatory networks correspond to master regulators [35]. Conversely, bottlenecks (nodes of high 'between-ness') often correspond to nodes that function as important information conduits, particularly in directed networks such as metabolic networks (in which metabolites flow between nodes) or signaling networks (in which information in the form of activations flows between nodes) [36]. Furthermore, these nodes most likely function as connectors between pathways, thereby mediating cross-talk (see later). Likewise, essential regulators can be identified by searching for composite hubs (i.e. nodes that have many interacting partners both in the metabolic and signaling network [37]). A combination of network stat-



(c)

Pathway	Number of node	Number of edges	Connected components	Number of core nodes	Number of extended nodes	Average degree (K)	Clustering coefficient (C)	Eccentricity (E)	Betweeness (B)
Notch	459	683	3	36	423	20.47	0.17	6.64	6202.85
Hedgehog	212	238	9	38	174	6.82	0.06	5.13	302.60

Figure 3. Differences in the topology of Notch and Hedgehog embedded signaling pathways. The Notch (**a**) and Hedgehog (**b**) networks were constructed from HPRD data and the corresponding statistics are listed (**c**). Blue nodes represent the core components of pathways and yellow nodes represent their interacting proteins. Although Notch and Hedgehog embedded pathways have a comparable number of core proteins (36 and 38, respectively), they have vastly different topologies (see Supplementary Table S6 for detailed definition of network topologies). The clustering coefficient (C), degree (K) and 'between-ness' (B). Values are all significantly higher in the Notch pathway (highlighted in red) than in the Hedgehog pathway. Note, K, C and B are statistical parameters used to measure network topologies (see Supplementary Table S2 for full definitions).

istics across different networks can be employed to identify several nodes that might have key roles in the biological function of cells. As knowledge of the different types of networks increases, an unambiguous and rigorous definition of protein function will eventually emerge from a combination of topological measures and network position [38]. That is, the importance of a protein is not only defined by its classical biochemical function, but also its position in the network. For example, hubs in Notch pathway include the histone deacetylases (HDAC), cAMP response element-binding protein (CREB)-binding protein (CREBBP), E1A binding protein p300 (EP300) and DVL2; all of these nodes have crucial roles in the regulation of the pathway. Furthermore, axis inhibitor 1 (AXIN1) is a bottleneck in both the Hedgehog and the Wnt pathways, indicating its importance for the information flow both within and between these pathways.

Examining crosstalk between embedded pathways

In living organisms, pathways are not isolated entities. From a systems biology perspective, pathways are linked together through crosstalk to perform biological functions as a system. In biology, the term 'crosstalk' refers to the phenomenon that signal components in signal transduction can be shared between different signaling pathways, and responses to a signal-inducing condition (e.g. stress) can activate multiple responses in the cell or organism. This crosstalk can be exemplified by protein kinase C, which is shared by the mitogen-activated protein kinase (MAPK), calcium, phosphatidylinositol, Wnt and vascular endothelial growth factor (VEGF) signaling pathways. However, because classical pathways only contain core components, they are insufficient to study crosstalk. This is evidenced by the few overlaps between classical pathways, which serve as indicators of the extent of crosstalk between them (Figure 4 and Supplementary Table S4). By contrast, embedded pathways provide an excellent platform to examine crosstalk because components of pathways are essentially embedded within a bigger network, which enables systematic identification of overlapping and linking components.

Figure 4 and Supplementary Table S4 show the overlaps between embedded pathways that correspond to signaling pathways in humans. The overlaps between both the core and the extended embedded pathways have been examined: the larger the overlap between embedded

(a)

Pathway	KEGG	BioCarta
MAPK	272	80
TGF-β	85	17
Notch	46	6
Wnt	148	25
Hedgehog	54	11
mTOR	50	22
TLR	92	36
Jak-STAT	154	5
VEGF	75	15

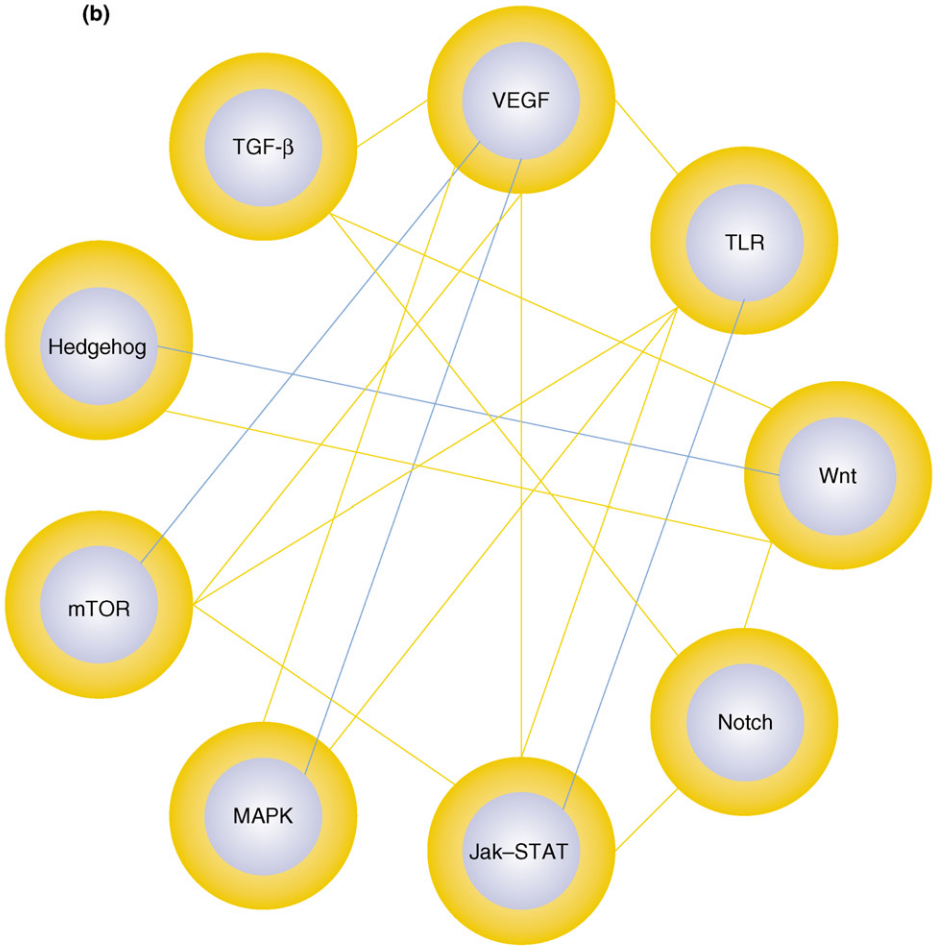


Figure 4. Overlaps between core and extended embedded pathways can provide insights into pathway crosstalk. (**a**) Discrepancies between the number of core components in KEGG and BioCarta are listed. (**b**) Blue nodes represent the core pathways, blue edges connect two core pathways that have significant overlaps (p-value ≤ 0.01). The yellow nodes and edges represent the extended pathways and the significant overlaps between extended pathways. [Part (b) uses the same color-coding as Figure 1.]

pathways, the more crosstalk takes place between the two pathways. Although most of the core embedded pathways do not overlap significantly, the corresponding extended embedded pathways often show a significant increase in overlap (Supplementary Table S4). These results indicate that many proteins exist as liaison components between pathways, and all the signaling pathways can be connected with just one degree of separation. A careful examination of these intermediate proteins might be useful in unraveling the mechanisms by which different pathways are related to each other.

Developing a simple version of edge ontology for pathways

As mentioned, classical pathway representations are often ambiguous because they use the same symbol to represent different functions. In the post-genomic era, this problem is further confounded by the emergence of various types of HTP data, which reveal different relationships between pathway components. In addition to protein–protein interaction networks, the core components of a pathway can also be mapped onto other types of networks, such as gene expression and regulatory networks. Simple edges (e.g. arrows) that are traditionally used in classical pathway representations might not be sufficient to meet the challenges of integrating these heterogeneous datasets. To perform large-scale mining of pathways, a precise edge ontology (or arrow ontology) must be developed to represent different types of relationships between pathway components.

To make things more complicated, many pathway databases are currently available [9,39]. Unfortunately, they do typically not share data models, file formats or access method. To foster sharing of these different information sources, several Extensible Markup Language (XML) exchange formats have been developed. System Biology Markup Language (SBML) [40] and CellML [41] focus mainly on quantitatively simulating concentrations of pathway components. The Proteomic Standards Initiative's Molecular Interaction (PSI-MI) [42] is an exchange format for molecular interaction, and the Biological Pathway Exchange (BioPAX) [43] is a more general format used to describe biological pathways.

Much effort has been devoted to developing consistent representations of pathways; however, most of these efforts focus on enumerating diverse types of edges. BioPAX has

been developing an ontology of interactions that reveals relationships between edges. To perform large-scale mining of pathways, making explicit the relationships between edges is an important step for elucidating the transitions and reactions between molecules. A precise edge (or arrow) ontology might also help improve pathway representation by highlighting both different types of relationships between pathway components and the knowledge of that relationship.

As an example, phosphorylation, ubiquitylation, glycosylation and methylation can all be viewed as types of reactions by transferring 'tags' to target proteins. Thus, an

Table 1. A prototype of an edge ontology

	Direction (level I)	Type (level II)	Sub-type (level III)	Specification (level IV)
Interaction	Directed	Tagging of proteins[a]	Phosphorylation	Serine[d]
				Tyrosine[d]
				Other[d]
			Dephosphorylation	
			Ubiquitylation	
			Glycosylation	N-linked
				O-linked
			Methylation	
		Cleavage of proteins[a]		
		Translocation	Diffusion	
			Active transport	
		Conformational change		
		Chemical reaction[b]		
		Catalysis		
		Unknown/other		
	Undirected	Binding[c]		
		Complex association[c]		
		Binding or association[c]		
		Dissociation		
		Co-expression		
		Unknown/other		

[a] describes 'inhibition' (where applicable).
[b] Double arrow describes reversible chemical reactions.
[c] 'Binding' describes direct physical interaction. 'Association' describes two proteins that are linked in the same complex but do not directly physically interact. 'Binding or association' describes the common scenario that arises in tandem affinity purification (TAP)-tagging experiments when the specific type of interaction is not known.
[d] Ser describes serine inhibition. A similarly annotated symbol can be used for tyrosine inhibition and so on.

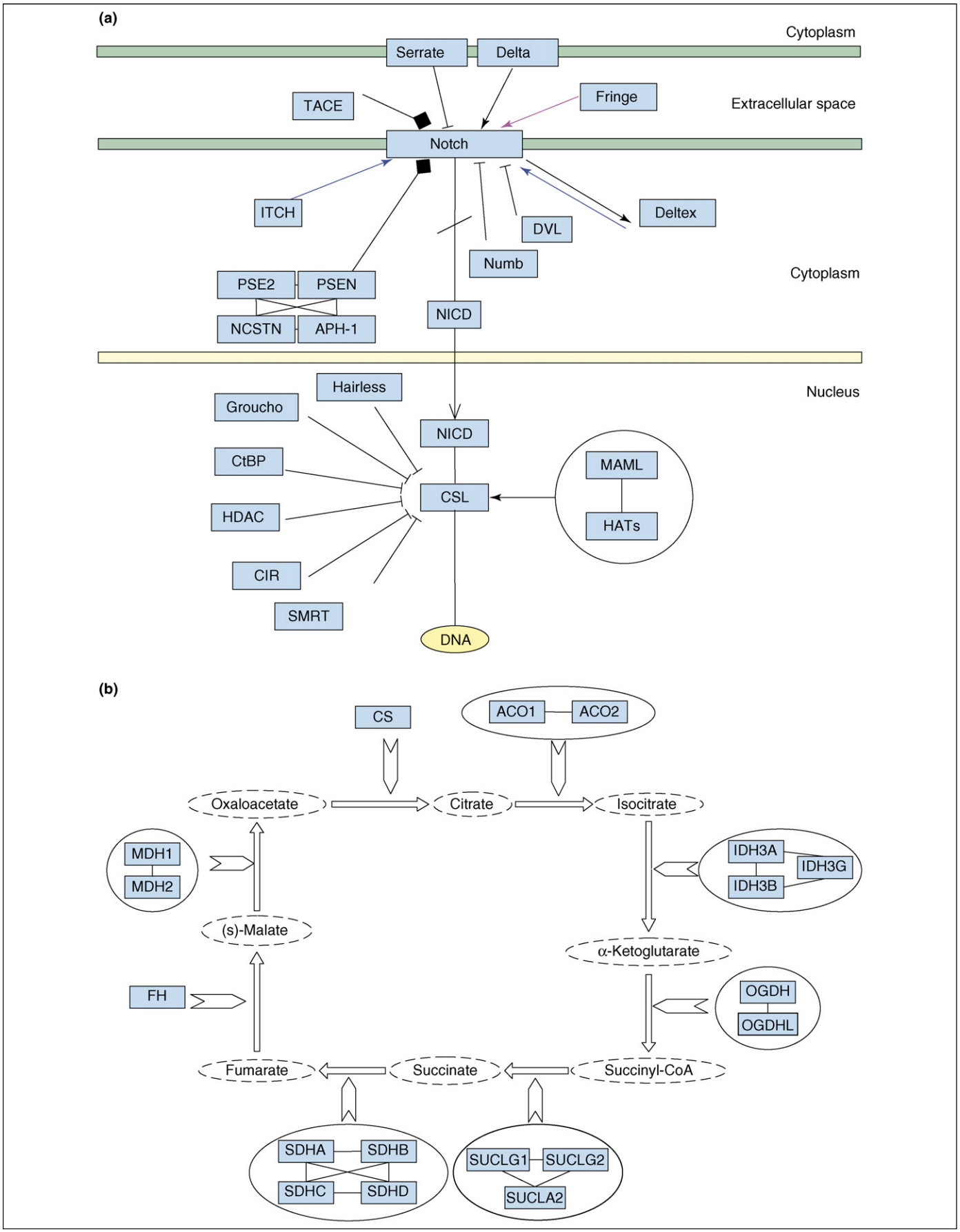


Figure 5. Biochemical pathways redrawn using the new edge ontology. We believe that pathways can be redrawn in a more representative manner using our proposed new edge ontology. Four different types of pathway are shown to highlight how more information can be conveyed using a more descriptive ontology. In particular, compare (a) with Figure 2a. In this case, for example, the interaction between TACE and Notch is more clearly seen to be a cleavage reaction in (a), whereas the type of interaction

ontology of edges that not only enumerates different types of edges but also classifies the edges into groups should be developed to capture this information. This explicit hierarchy of relationships can be exploited to enable accurate computational analysis without losing the expressiveness of the classical representation. For example, some pathway interactions can be represented by specific symbols, such as serine phosphorylation. However, translation to a more general interaction, such as 'tagging' that leads to activation, can be employed to perform high-level analysis of the pathway or to compare multiple pathways represented at a different level of specification.

Here, we propose a simple version of edge ontology to illustrate how we could deal with this issue in the future. This ontology provides both an unambiguous definition of the interactions and defines a hierarchy of those interactions. In particular, the hierarchy of interactions might be useful to obtain multiple views of a pathway, from a general one to a more specific one. It also contains symbols that might help the graphical representation of interactions. Please note, this ontology is far from complete. We foresee that the formidable goal of constructing a complete ontology for pathways would take multiple groups many years to achieve. However, this simple ontology could be used as a starting point from which we hope a complete ontology can be built. We also realize that a consistent representation of nodes is equally important; however, this problem can be largely solved by using Gene Ontology (GO) [44]. The GO provides a controlled vocabulary for describing gene and gene-product attributes in any organism, and hence could be used as an approximate node ontology. Here, we focus on edge ontology.

The edge ontology we propose is presented in Table 1. We use different shapes, symbols and colors to represent diverse types of interactions between pathway components. We also define a simple hierarchy of interactions from general ones to more specific ones. The first level divides directed from undirected interactions, whereas the second level highlights the main mechanisms of interaction, which are, in turn, defined in more detail in the third level. The fourth level further specifies some of the interaction types. The edges in the second level have different shapes, whereas those in the third level are represented by different colors. Further specifications can be defined by adding annotations on the edge, such as those in the fourth level. Nearly all the edges connect two components of the pathways, such as proteins and molecules, except that the 'catalysis' edge connects a pathway component to a 'chemical reaction' edge. This enables us to properly describe metabolic pathways that typically display a sequence of chemical reactions in which enzymes take part.

For example, a black arrow is used to indicate a 'tagging' interaction, meaning an interaction that binds a molecule to a pathway component. If we know the type of interaction in more detail, different colors can be used to describe different 'tagging' mechanisms: red for phosphorylation, blue for ubiquitylation and so on. In many cases, however, the tagging mechanism activates proteins; to highlight that some tagging mechanisms inhibit proteins, a solid vertical line is used instead of an arrowhead. To add further detail to the relationship, an annotation on the edge can be used, such as 'ser' for serine phosphorylation and 'N' for N-linked glycosylation. It is worth noting that symbols from different levels can be used concurrently in the same pathway. This might be used to emphasize the level of understanding of the interaction.

For illustration purposes only, we provide four examples of the application of this edge ontology: the Notch pathway, the citric acid cycle, the JAK–STAT signaling pathway and the caspase cascade pathway. Figure 5 shows the classical pathways redrawn according to our new edge ontology.

Concluding remarks

Although classical representations of biochemical pathways can provide in-depth views of isolated sets of genes, the network approach is capable of analyzing pathways on three different levels: whole system (crosstalk), whole network and individual nodes. Whereas embedding pathways to large-scale protein–protein interaction networks enables easy comparison of properties across, between and within pathways, we also experience substantial information loss. One way to circumvent this problem would be to overlay additional types of HTP data onto the pathway by defining different types of edges.

To properly analyze this type of multilayered network, a precise edge ontology must be defined. The edge ontology should provide an unambiguous representation of the relationships between biomolecules in addition to revealing relationships between the edges. However, even a well-defined edge ontology suffers the limitation of lacking explicit temporal information. Properly incorporating explicit temporal information will be the next challenge in the representation of pathways.

between the two components is not depicted in Figure 2a. Note, solid ellipses have been used to enclose complexes, and the complexes are also linked by solid edges to indicate known physical interactions. In other contexts, different edges can be used given the uncertainty over the types of actual association. (**a**) In the Notch signaling pathway, Fringe activates Notch by glycosylation (pink solid arrow). Delta activates (black solid arrow) Notch and Serrate inhibits (black with bar) Notch. TACE catalyzes the cleavage (black with solid diamond) of Notch. Binding edges (black lines) are used consistently for the components of the PSE2–PSEN–NCSTN–APH-1 complex. NICD translocates (black open arrow) into the nucleus and promotes transcription in combination with CSL, MAML and HATs. Abbreviations: APH-1, anterior pharynx defective 1 homolog A; CIR, CBF1-interacting co-repressor; CtBP, C-terminal-binding protein; CSL, recombining binding protein suppressor of hairless; Delta, delta-like 3; Deltex, deltex homolog 2; Fringe, LFNG O-fucosylpeptide 3-b-N-acetylglucosaminyltransferase; HATS, histone acetyltransferases; HDAC, histone deacetylase; IICH, itchy homolog E3 ubiquitin protein ligase; MAML, Mastermind; NCSTN, nicastrin; NICD, Notch Intra-cellular domain; Notch, Notch homolog 1, translocation-associated; NUMB, numb homolog; PSE2, presenilin enhancer 2 homolog; PSEN, presenilin 1; Serrate, jagged 1; SMRT, nuclear receptor co-repressor 2. (**b**) In the citric acid cycle, the 'catalysis' edge connects an enzyme to a 'chemical reaction' edge (see Table 1 for key). The main chemicals that take part in the interactions are shown in broken ellipses. We found direct interaction between IDH3A, IDH3B and IDH3G; however, the specific type of interaction involved in this complex, such as that from tandem affinity purification (TAP)-tagging experiments, is not known. Thus, we could use 'association' or 'binding or association' defined in the edge ontology to represent the pair-wise relationship between proteins in this complex. Abbreviations: ACO1, aconitase 1; ACO2L, aconitase 2; CS, citrate synthase; FH, fumarate hydratase; IDH, isocitrate dehydrogenase; MDH, malate dehydrogenase; OGDH, oxoglutarate (a-ketoglutarate) dehydrogenase (lipoamide); OGDHL, oxoglutarate dehydrogenase-like; SDHA, SDHB, SDHC and SDHD, succinate dehydrogenase complex, subunit A, B, C and D, respectively; SUCLA2, succinate-CoA ligase (ADP-forming), β subunit; SUCLG1, succinate-CoA ligase (GDP-forming), α subunit; SUCLG2, succinate-CoA ligase (GDP-forming), β subunit.

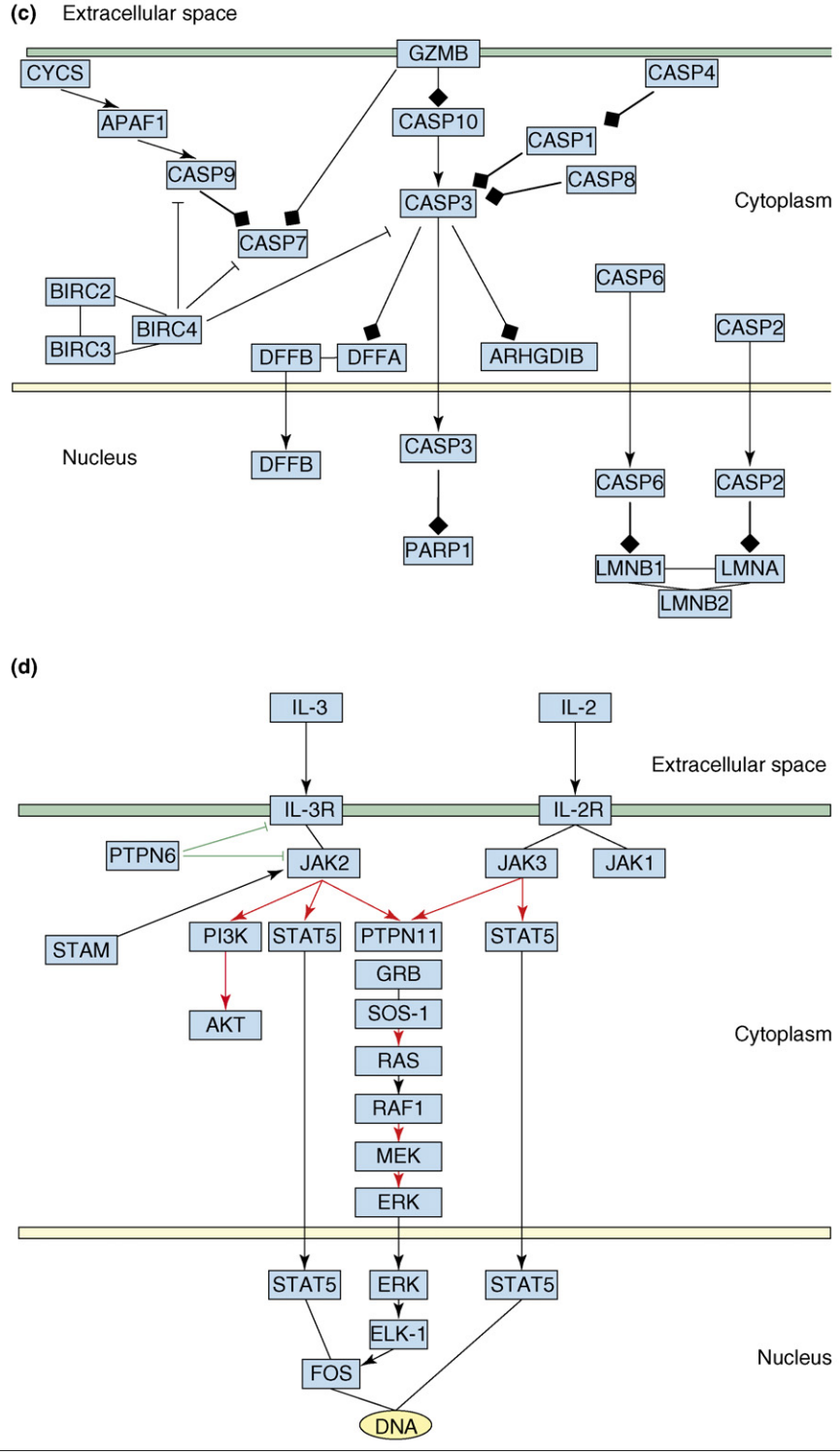


Fig. 5. Cont. (**c**) The caspase cascade involves mainly cleavage interactions among caspases, which cleave proteins after an aspartic acid residue. Abbreviations: APAF1, apoptotic peptidase activating factor 1; ARHGDIB, Rho GDP dissociation inhibitor (GDI) β; BIRC, baculoviral IAP repeat-containing; CASP, caspase; CYCS, cytochrome c, somatic; DFFA, DNA fragmentation factor (DFF), 45-kDa, α polypeptide; DFFB, DFF, 40-kDa, β polypeptide; GZMB, granzyme B; LMNA, lamin A/C; LMNB, lamin B; PARP1, poly (ADP-ribose) polymerase-1. (**d**) A portion of the JAK–STAT signaling pathway, including the response to IL-2 and IL-3. JAKs bind to interleukin receptors and are activated by the binding of the ligand (black solid arrow). JAK1 and JAK3 activate STAT5 by phosphorylation (red solid arrow), which translocates (black open arrow) to the nucleus and activates transcription of its target genes (black line). PTPN6 inhibits the cytokine receptor and JAK1 through dephosphorylation (green with bar). JAK1 and JAK3 activate PTPN11, which is bound to GRB and SOS-1. This PTPN11–GRB–SOS-1 complex activates the MAPK signaling pathway. Abbreviations: AKT, v-akt murine thymoma viral oncogene homolog 3 (protein kinase B,γ); ELK-1, member of ETS oncogene family; ERK, mitogen-activated protein kinase 3; FOS, v-fos FBJ murine osteosarcoma viral oncogene homolog; GRB, growth factor receptor-bound protein 2; IL, interleukin; IL-3R, interleukin 3 receptor; JAK, Janus kinase; MEK, mitogen-activated protein kinase kinase 1; PI3K, phosphatidylinositol 3-kinase; PTPN, protein-tyrosine phosphatase; RAF1, v-raf-1 murine leukemia viral oncogene homolog 1; RAS, oncogene homolog 2; SOS-1, son of sevenless homolog 1; STAM, signal-transducing adapter molecule; STAT, signal transducer and activator of transcription.

Acknowledgements
This work is supported by an NIH grant to M.B.G. We thank Ashish Agarwal and Emmett Sprecher for valuable comments on improving this manuscript.

Supplementary data

Supplementary material associated with this article can be found online at doi:10.1016/j.tibs.2007.06.003.

References

1 Keseler, I.M. *et al.* (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.* 33 (Database issue), D334–D337
2 Hackl, H. *et al.* (2004) GOLD.db: genomics of lipid-associated disorders database. *BMC Genomics* 5, 93
3 Kanehisa, M. *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32 (Database issue), D277–D280
4 Overbeek, R. *et al.* (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* 28, 123–125
5 Karp, P.D. *et al.* (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* 33, 6083–6089
6 Caspi, R. *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 34 (Database issue), D511–D516
7 Dahlquist, K.D. *et al.* (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* 31, 19–20
8 Joshi-Tope, G. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 33 (Database issue), D428–D432
9 Cary, M.P. *et al.* (2005) Pathway information for systems biology. *FEBS Lett.* 579, 1815–1820
10 Schaefer, C.F. (2004) Pathway databases. *Ann. N. Y. Acad. Sci.* 1020, 77–91
11 Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
12 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574
13 Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
14 Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183
15 Gavin, A.C. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631–636
16 Li, S. *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* 303, 540–543
17 Giot, L. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science* 302, 1727–1736
18 Rual, J.F. *et al.* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437, 1173–1178
19 Han, J.D. *et al.* (2005) Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.* 23, 839–844
20 Suthram, S. *et al.* (2005) The *Plasmodium* protein network diverges from those of other eukaryotes. *Nature* 438, 108–112
21 Rain, J.C. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature* 409, 211–215
22 Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113
23 Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature* 393, 440–442
24 Fraser, H.B. *et al.* (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750–752
25 Jordan, I.K. *et al.* (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 12, 962–968
26 Kim, P.M. *et al.* (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 314, 1938–1941
27 Luscombe, N.M. *et al.* (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431, 308–312
28 Wittig, U. and De Beuckelaer, A. (2001) Analysis and comparison of metabolic pathway databases. *Brief. Bioinform.* 2, 126–142
29 Peri, S. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 13, 2363–2371
30 Kadesch, T. (2004) Notch signaling: the demise of elegant simplicity. *Curr. Opin. Genet. Dev.* 14, 506–512
31 Bray, S.J. (2006) Notch signalling: a simple pathway becomes complex. *Nat. Rev. Mol. Cell Biol.* 7, 678–689
32 Salinas, P.C. (2005) Retrograde signalling at the synapse: a role for Wnt proteins. *Biochem. Soc. Trans.* 33, 1295–1298
33 Espinosa, L. *et al.* (2003) Phosphorylation by glycogen synthase kinase-3β down-regulates Notch activity, a link for Notch and Wnt pathways. *J. Biol. Chem.* 278, 32227–32235
34 Lleo, A. *et al.* (2003) Notch1 competes with the amyloid precursor protein for γ-secretase and down-regulates presenilin-1 gene expression. *J. Biol. Chem.* 278, 47370–47375
35 Borneman, A.R. *et al.* (2006) Target hub proteins serve as master regulators of development in yeast. *Genes Dev.* 20, 435–448
36 Yu, H. *et al.* (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* 3, e59
37 Yu, H. *et al.* (2006) Design principles of molecular networks revealed by global comparisons and composite motifs. *Genome Biol.* 7, R55
38 Huynen, M.A. *et al.* (2003) Function prediction and protein networks. *Curr. Opin. Cell Biol.* 15, 191–198
39 Bader, G.D. *et al.* (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.* 34 (Database issue), D504–D506
40 Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531
41 Lloyd, C.M. *et al.* (2004) CellML: its future, present and past. *Prog. Biophys. Mol. Biol.* 85, 433–450
42 Hermjakob, H. *et al.* (2004) The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. *Nat. Biotechnol.* 22, 177–183
43 Stromback, L. and Lambrix, P. (2005) Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics* 21, 4401–4407
44 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29