The BioPAX community standard for pathway data sharing

Emek Demir[1,2,*], Michael P Cary[1], Suzanne Paley[3], Ken Fukuda[4], Christian Lemer[5], Imre Vastrik[6], Guanming Wu[7], Peter D'Eustachio[8], Carl Schaefer[9], Joanne Luciano[10], Frank Schacherer[11], Irma Martinez-Flores[12], Zhenjun Hu[13], Veronica Jimenez-Jacinto[12], Geeta Joshi-Tope[14], Kumaran Kandasamy[15], Alejandra C Lopez-Fuentes[16], Huaiyu Mi[17], Elgar Pichler[18], Igor Rodchenkov[19], Andrea Splendiani[20,21], Sasha Tkachev[22], Jeremy Zucker[23], Gopal Gopinath[24], Harsha Rajasimha[25,26], Ranjani Ramakrishnan[27], Imran Shah[28], Mustafa Syed[29], Nadia Anwar[1], Özgün Babur[1,2], Michael Blinov[30], Erik Brauner[31], Dan Corwin[32], Sylva Donaldson[19], Frank Gibbons[31], Robert Goldberg[33], Peter Hornbeck[22], Augustin Luna[34], Peter Murray-Rust[35], Eric Neumann[36], Oliver Reubenacker[37], Matthias Samwald[38,39], Martijn van Iersel[40], Sarala Wimalaratne[41], Keith Allen[42], Burk Braun[11], Michelle Whirl-Carrillo[43], Kei-Hoi Cheung[44], Kam Dahlquist[45], Andrew Finney[46], Marc Gillespie[47], Elizabeth Glass[29], Li Gong[43], Robin Haw[7], Michael Honig[48], Olivier Hubaut[5], David Kane[49], Shiva Krupa[50], Martina Kutmon[51], Julie Leonard[42], Debbie Marks[52], David Merberg[53], Victoria Petri[54], Alex Pico[55], Dean Ravenscroft[56], Liya Ren[14], Nigam Shah[57], Margot Sunshine[34], Rebecca Tang[43], Ryan Whaley[43], Stan Letovksy[58], Kenneth H Buetow[59], Andrey Rzhetsky[60], Vincent Schachter[61], Bruno S Sobral[25], Ugur Dogrusoz[2], Shannon McWeeney[27], Mirit Aladjem[34], Ewan Birney[6], Julio Collado-Vides[12], Susumu Goto[62], Michael Hucka[63], Nicolas Le Novère[6], Natalia Maltsev[29], Akhilesh Pandey[15], Paul Thomas[17], Edgar Wingender[64], Peter D Karp[3], Chris Sander[1] & Gary D Bader[19]

Biological Pathway Exchange (BioPAX) is a standard language to represent biological pathways at the molecular and cellular level and to facilitate the exchange of pathway data. The rapid growth of the volume of pathway data has spurred the development of databases and computational tools to aid interpretation; however, use of these data is hampered by the current fragmentation of pathway information across many databases with incompatible formats. BioPAX, which was created through a community process, solves this problem by making pathway data substantially easier to collect, index, interpret and share. BioPAX can represent metabolic and signaling pathways, molecular and genetic interactions and gene regulation networks. Using BioPAX, millions of interactions, organized into thousands of pathways, from many organisms are available from a growing number of databases. This large amount of pathway data in a computable form will support visualization, analysis and biological discovery.

Increasingly powerful technologies, including genome-wide molecular measurements, have accelerated progress toward a complete map of molecular interaction networks in cells and between cells of many organisms. The growing scale of these maps requires their representation in a form suitable for computer processing, storage and dissemination by means of software systems. The BioPAX project aims to facilitate knowledge representation, systematic collection, integration and wide distribution of pathway data from heterogeneous information sources. This will enable these data to be incorporated into distributed biological information systems that support visualization and analysis.

BioPAX supports efforts working toward a complete representation of basic cellular processes. Biology has come a long way since the Boehringer-Mannheim wall chart of metabolic pathways[1] and the Nicholson Metabolic Map[2]. Since then, several groups have developed methods and databases for organizing pathway information[3–16], but only recently have groups collaborated as part of the BioPAX project to develop a generally accepted standard way of representing these pathway maps. Complete molecular process maps must include all interactions, reactions, dependencies, influence and information flow between pools of molecules in cells and between cells. For ease of use and simplicity of presentation, such network maps are often organized in terms of subnetworks or pathways. Pathways are models delineated within the entire cellular biochemical network that help us describe and understand specific biological processes. Thus, a useful definition of a pathway is a set of interactions between physical or genetic cell components, often describing a cause-and-effect or time-dependent process, that explains observable biological phenomena. How do we represent these pathways in a generally accepted and computable form?

Challenges posed by the many fragmented pathway databases

The total volume of pathway data mapped by biologists and stored in databases has entered a rapid growth phase, with the number of

*A full list of author affiliations appear at the end of this paper.

Published online 9 September 2010; doi:10.1038/nbt.1666

online resources for pathways and molecular interactions increasing 70%, from 190 in 2006 to 325 in 2010 (ref. 17). In addition, molecular profiling methods, such as RNA profiling using microarrays, or protein quantification using mass spectrometry, provide large amounts of information about the dynamics of cellular pathway components and increase the power of pathway analysis techniques[18,19]. However, this growth poses a formidable challenge for pathway data collection and curation as well as for database, visualization and analysis software, as these data are often fragmented.

The principal motivation for building pathway databases and software tools is to facilitate qualitative and quantitative analysis and modeling of large biological systems using a computational approach. Over 300 pathway or molecular interaction–related data resources[17] and many visualization and analysis software tools[3,20–22] have been developed. Unfortunately, most of these databases and tools were originally developed to use their own pathway representation language, resulting in a heterogeneous set of resources that are extremely difficult to combine and use. This has occurred because many different research groups, each with their own system for representing biomolecules and their interactions in a pathway, work independently to collect pathway data recorded in the literature (estimated from text-mining projects[23] to be present in at least 10% of the >20 million articles currently indexed by PubMed). As a result, researchers waste time collecting information from different sources and converting it from one form of representation to another. Fragmented pathway data results in substantial lost opportunity cost. For instance, visualization and analysis tools developed for one pathway database cannot be reused for others, making software development efforts more expensive. Therefore, it is imperative to develop computational methods to cope with both the magnitude and fragmented nature of this expanding, valuable pathway information. Whereas independent research efforts are needed to find the best ways to represent pathways, community coordination and agreement on standard semantics is necessary to be able to efficiently integrate pathway data from multiple sources on a large scale.

BioPAX requirements and implementation

A common, inclusive and computable pathway data language is necessary to share knowledge about pathway maps and to facilitate integration and use for hypothesis testing in biology[24]. A shared language facilitates communication by reducing the number of translations required to exchange data between multiple sources (**Fig. 1**). Developing such a representation is challenging owing to the variety of pathways in biology and the diverse uses of pathway information. Pathway representations frequently use abstractions for metabolic, signaling, gene regulation, protein interaction and genetic interaction, and these serve as a starting point toward a shared language[25]. Also, several variants of this common language may be required to answer relevant research questions in distinct fields of biology, each covering unique levels of detail addressing different uses, but these should be rooted in common principles and must remain compatible.

BioPAX addresses these challenges. We developed BioPAX as a shared language to facilitate communication between diverse software systems and to establish standard knowledge representation of pathway information. BioPAX supports representation of metabolic and signaling pathways, molecular and genetic interactions and gene regulation. Relationships between genes, small molecules, complexes and their states (e.g., post-translational protein modifications, mRNA splice variants, cellular location) are described, including the results of events. Details about the BioPAX language are available in online documentation at http://www.biopax.org/. The BioPAX language

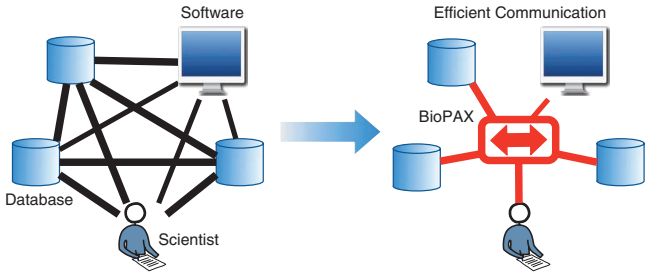

Figure 1 BioPAX is a shared language for biological pathways. BioPAX reduces the effort required to efficiently communicate between pathway users, databases and software tools. Without a shared language, each system must speak the language of all other systems in the worst case (black lines). With a shared language, each system only needs to speak that language (central red box).

provides terms and descriptions, to represent many aspects of biological pathways and their annotation. It is implemented as an ontology, a formal system of describing knowledge (**Box 1**) that helps structure pathway data so that they are more easily processed by computer software (**Fig. 2**). It provides a standard syntax used for data exchange that is based on OWL (Web Ontology Language) (**Box 1**). Finally, it provides a validator that uses a set of rules to verify whether a BioPAX document is complete, consistent and free of common errors. BioPAX is the only community standard for biological pathway exchange to and from databases, but it is related to other standards (discussed below in the "What is not covered?" section).

Example of a pathway in BioPAX

Pathway models are generally described with text and with network diagrams. Here we use the AKT signaling pathway[26,27] as an example to show how a typical pathway diagram that can only be interpreted by people (**Fig. 3**, top left) would be represented using BioPAX (**Fig. 3**, right). The AKT pathway is a cell surface receptor–activated signaling cascade that transduces external signals to intracellular events through a series of steps including protein-protein interactions and protein kinase–mediated phosphorylation. The pathway eventually activates transcription factors, which turn on genes to promote cell survival. By representing the pathway using the BioPAX language (**Fig. 3** and **Supplementary Tables 1** and **2**), it can be analyzed by computational approaches, such as pathway analysis of gene expression data.

Representing a pathway using the BioPAX language sometimes necessitates being more explicit to avoid capturing inconsistent data. For instance, the typical notion of an 'active protein' is dependent on context, as the same molecule could be active in one cellular context, such as a cellular compartment with a set of potentially interacting molecules, and inactive in another context. Thus, capturing the specific mechanism of activation, such as phosphorylation modification, is usually required, and the presence of downstream events that include the modified form signifies that the molecule is active. Interactions where the mechanism of action is unknown can also be specified.

What does BioPAX include?

BioPAX covers all major concepts familiar to biologists studying pathways, including metabolic and signaling pathways, gene regulatory networks and genetic and molecular interactions (**Supplementary Table 3**). The BioPAX language is distributed as an ontology definition (**Fig. 4**) with associated documentation, a validator for checking a BioPAX document for errors and other software tools (**Table 1**).

Box 1 What is an ontology?

An ontology is a formal system for representing knowledge[64]. Such representation is required for computer software to make use of information. Example ontologies include organism taxonomies[65] and the Gene Ontology[40]. A formal representation allows consistent communication of knowledge among individuals or computer systems and helps manage complexity in information processing as knowledge is broken down into clear concepts that can be considered independently. Ontologies also enable integration of knowledge between independent resources linked on the World Wide Web. Such linked, structured data form the basis of the semantic web, an extension of the web that promises improved information management and search capability[61]. Representing and sharing knowledge using ontologies is simplified by availability of the standard web ontology language (OWL; http://www.w3.org/TR/owl-features/). Tools to edit OWL, such as Protégé[63], have been developed by the semantic web community and adopted in the life sciences. Implementing BioPAX using OWL enables both the ontology and the individuals and values to be stored in the same XML-based format, which makes data transmission easier. Using OWL also enables BioPAX users to take advantage of existing software tools for editing, transmitting, querying, reasoning about and visualizing OWL data.

An ontology is composed of classes, properties (representing relations) and restrictions and is used to define individuals (instances of classes, also known as objects) and values for their properties. Classes (also known as concepts or types) are often arranged into a hierarchy (or taxonomy) where child classes are more specific than, and inherit the properties of, parent classes. For example, in BioPAX, the BiochemicalReaction class is a subclass of the Conversion class. Classes may have properties (also known as fields, attributes or slots), which express possible relations to other classes (that is, they may have values of specific types). For example, a SmallMolecule is related to the ChemicalStructure class by the property structure. Restrictions (also known as constraints) define allowable values and connections within an ontology. For example, molecularWeight must be a positive number. Individuals are instances of classes where values occupy the properties of those instances. BioPAX defines the classes, properties and restrictions required to represent biological pathways and leaves creation of the individuals to users (data providers and consumers).

Pathway abstractions frequently used in several pathway databases and software programs are supported as follows:

• Metabolic pathways are described using the 'enzyme, substrate, product' abstraction[28] where substrates and products of a biochemical reaction are often small molecules. An enzyme, often a protein, catalyzes the reaction, and inhibitors and activators can modulate the catalysis event. Metabolic pathways use BioPAX classes: PhysicalEntity, Conversion, Catalysis, Modulation, Pathway.

• Signaling pathways involve molecules and complexes participating in biochemical reactions, binding, transportation and catalysis events (**Fig. 3**)[5,9,29–31]. These pathways may also include descriptions of molecular states (such as cellular location, covalent and noncovalent modifications, as well as fragments of sequence cleaved from a precursor) and generic molecules (such as the family of homologous Wnt proteins). Signaling pathways use BioPAX classes: PhysicalEntity, Conversion, Control, Catalysis, Modulation, MolecularInteraction, Pathway.

• Gene regulatory networks involve transcription and translation events and their control[12,14]. Transcription, translation and other template-directed reactions involving DNA or RNA are captured in a 'template reaction' in BioPAX, which maps a template to its encoded products (e.g., DNA to mRNA). Multiple sequence regions on a single strand of the template, such as promoters, terminators, open reading frames, operons and various reaction machinery binding sites, are active in a template reaction. Transcription factors (generally proteins and complexes), microRNAs and other molecules, participate in a 'template reaction regulation' event. Gene regulatory networks use BioPAX classes: PhysicalEntity, TemplateReaction, TemplateReactionRegulation.

• Molecular interactions, notably protein-protein[32–36] and protein-DNA interactions[37], involve two or more 'physical entities'. BioPAX follows the standard representation scheme of the Proteomics Standards Initiative Molecular Interaction (PSI-MI) format[38]. Molecular interactions use BioPAX classes: PhysicalEntity, MolecularInteraction.

• Genetic interactions occur between two genes when the phenotypic consequence of perturbing both genes is different than expected given the phenotypes of each single gene perturbation[39]. BioPAX represents this as a pair of genes that participate in a 'genetic interaction' measured using an observed 'phenotype'. Genetic interactions use BioPAX classes: Gene, GeneticInteraction.

Metabolic-, signaling- and gene regulatory–pathway abstractions are process oriented. They imply a temporal order and can be thought of as extensions of the standard chemical reaction pathway notation to accommodate biological information. Molecular and genetic interactions, however, imply a static network of connections among system components, instead of the temporally ordered process of reactions that defines a metabolic or signaling pathway. BioPAX supports combining these different types of data into a single model that is useful to gain a more complete view of a cellular process.

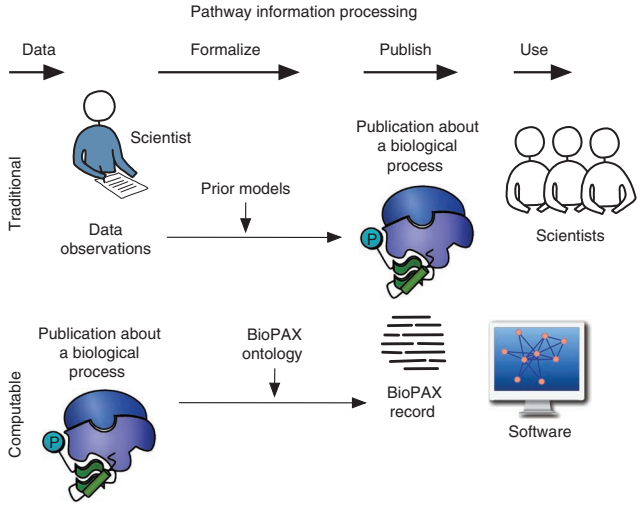

Figure 2 BioPAX enables computational data gathering, publication and use of information about biological processes. Traditional pathway information processing: observations considering prior models published as text and figures. Computable pathway information processing: scientist's description represented using formal, computable framework (ontology) published in a format readable by computer software for analysis by scientists.

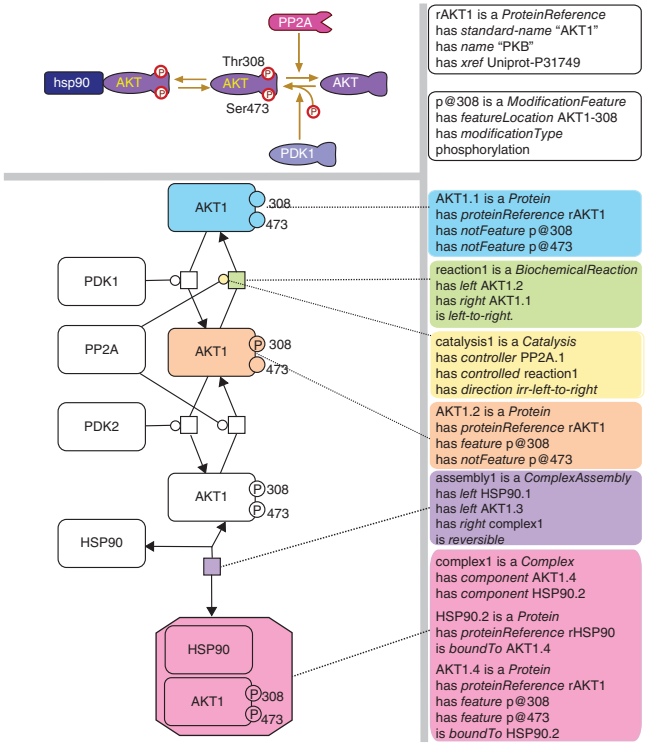

Figure 3 The AKT pathway as represented by a traditional method (top left; from http://www.biocarta.com/), a formalized SBGN diagram (left; from http://www.sbgn.org/[62]) and using the BioPAX language (right). An important advantage of the BioPAX representation is that it can be interpreted by computer software and used in multiple ways, including automatic diagram creation, information retrieval and analysis. Online documentation at http://www.biopax.org/ contains more details about how to represent diverse types of biological pathways. Actual samples of pathway data in BioPAX OWL XML format are available in **Supplementary Tables 1** and **2**.

BioPAX provides many additional constructs, not shown in **Figure 4**, that are used to store extra details, such as database cross-references, chemical structure, experimental forms of molecules, sequence feature locations and links to controlled vocabulary terms in other ontologies (**Supplementary Fig. 1**). BioPAX reuses a number of standard controlled vocabularies defined by other groups. For example, Gene Ontology[40] is used to describe cellular location, PSI-MI vocabularies[38] are used to define evidence codes, experimental forms, interaction types, relationship types and sequence modifications, and Sequence Ontology[41] is used to define types of sequence regions, such as a promoter region on DNA involved in transcription of a gene. Other useful controlled vocabularies can be referenced, such as the molecule role ontology[42].

BioPAX defines additional semantics that are currently only captured in documentation. For instance, physical entities represent pools of molecules and not individual molecules, corresponding to typical semantics used when describing pathways in textbooks or databases. A molecular pool is a set of molecules in a bounded area of the cell, thus it has a concentration. Pools can be heterogeneous and can overlap, as in the case of a protein existing in multiple phosphorylation states.

BioPAX also defines a range of constructs that are represented as ontology classes. Some of these represent biological entities, such as proteins, and are organized into classes that conceptualize the pathway knowledge domain. Others are used to represent annotations and properties of the database representation of biological entities. For instance, BioPAX provides 'xref' classes to represent different kinds of references to databases that can be useful for data integration. These are represented as subclasses of UtilityClass for convenience. A future version of BioPAX would ideally capture these semantics and structure these concepts more formally.

Uses of pathway data encoded in BioPAX

Once pathway data are translated into a standard computable language, such as BioPAX, it is easier for software to access them and thereby support browsing, retrieval, visualization and analysis (**Fig. 5**). This enables efficient reuse of data in different ways, avoiding the time-consuming and often frustrating task of translating them between formats (**Fig. 1**). Additionally, it enables uses that would be impractical without a standard format, such as those dependent on combining all available pathway data.

BioPAX can be used to help aggregate large pathway data sets by reducing the required collection and translation effort, for instance using software such as cPath[43]. Typical biological queries, such as 'What reactions involve my protein of interest?' generate more complete answers when querying these larger pathway data sets. Another frequent use is to find pathways that are active in a particular biological context, such as a cell state determined by a genome-scale molecular profile measurement. For instance, pathways with multiple differentially expressed genes may be transcriptionally active in one biological condition and not in another. Functional genomics and pathway data can be imported into software and combined for visualization and analysis to find interesting network regions. A typical workflow involves overlaying molecular profiling data, such as mRNA transcript profiles, on a network of interacting proteins to identify transcriptionally active network regions, which may represent active pathways[44]. A number of recent papers have used this pathway analysis workflow to highlight genes and pathways that are active in specific model organisms or diseased tissues, such as breast cancer, using gene and protein expression, copy number variants and single-nucleotide polymorphisms[19,44–49]. BioPAX has also been used in a number of these studies to collect and integrate large amounts of pathway information from multiple databases for analysis. For instance, protein expression data were combined with pathway information to highlight the importance of apoptosis in a mouse model of heart disease[50]. Multiple groups have found that tumor-associated mutations are significantly related by pathway

Table 1 What is included in BioPAX

Content	Description
Ontology specification	Web Ontology Language (OWL) XML file, developed using free Protégé ontology editor software[63].
Language documentation	Explanation of BioPAX entities, example documentation, best practice recommendations, use cases and instructions for carrying out frequently used technical tasks.
Example files	Example files for biochemical pathway, protein and genetic interaction, protein phosphorylation, insulin maturation, gene regulation and generic molecules in OWL XML.
Graphical representation	Recommendations for graphical representation using Systems Biology Graphical Notation (SBGN) as a guide.
Paxtools software	Java programming library supporting import/export, conversion and validation. Can be used to add BioPAX support to software.
List of data sources and supporting software	Databases making data available in BioPAX format, software systems for storing, visualizing and analyzing BioPAX pathways.

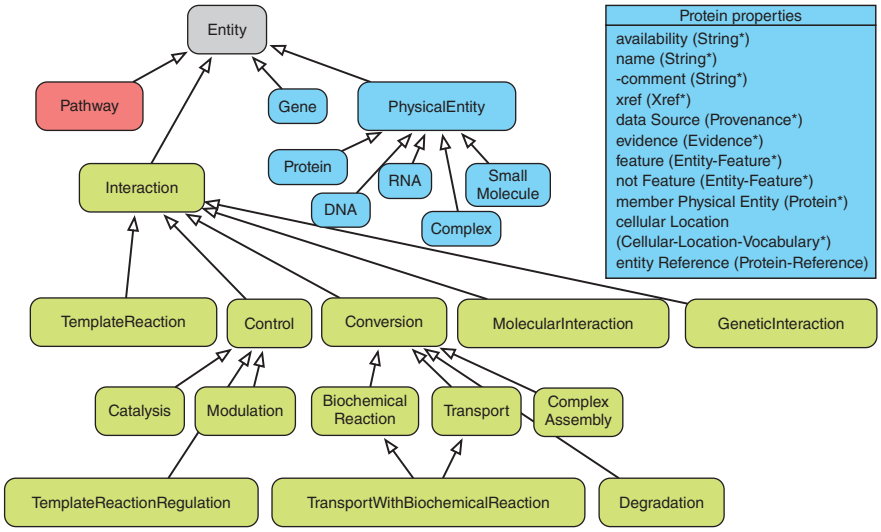

Figure 4 High-level view of the BioPAX ontology. Classes, shown as boxes and arrows, represent inheritance relationships. The three main types of classes in BioPAX are Pathway (red), Interaction (green) and PhysicalEntity and Gene (blue). For brevity, the properties of the Protein class only are shown as an example at the top right. Asterisks indicate that multiple values for the property are allowed. Refer to BioPAX documentation at http://www.biopax.org/ for full details of all classes and properties.

information[47,48]. And recently, in a study of rare copy number variants in 996 individuals with autism spectrum disorder, a core set of neuronal development–related pathways were found to link dozens of rare mutations to autism that were not significantly linked to the disorder on their own by traditional single-gene association statistics[49]. These studies highlight the importance of pathway information in explaining the functional consequence of mutations in human disease. BioPAX pathway data can also be converted into simulation models, for instance using differential equations[51] or rule-based modeling languages[52], to predict how a biological system may function after a gene is knocked out.

BioPAX is useful for exchanging information among and between data providers and analysis software. Pathway database groups can share the effort of pathway curation by making their pathways available in BioPAX format and exchanging them with others. For example, pathways in BioPAX format from the Reactome[8] database are imported by the US National Cancer Institute/Nature Pathway Information Database[9]. Data providers can use existing BioPAX-enabled software to add useful new features to their systems. For example, the Cytoscape network visualization software[20] can read and display BioPAX-formatted data as a network. The Reactome group used this feature to create a pathway visualization tool for their website. Because Reactome data were available in BioPAX format, and Cytoscape could already read BioPAX format, this new feature was easy to implement.

The Paxtools Java programming library for BioPAX has been developed to help software developers readily support the import, export and validation of BioPAX-formatted data for various uses in their software (http://www.biopax.org/paxtools/). Using Paxtools and other tools, a range of BioPAX-compatible software has been developed, including browsers, visualizers, querying engines, editors and converters (**Supplementary Table 4**). For instance, the ChiBE and VisANT pathway-visualization tools read BioPAX format[22], and the WikiPathways website[53], a community wiki for pathways, is working on using BioPAX to help import pathways from several sources, including manually edited pathways from biologists. The Pathway Tools software[21] and CellDesigner pathway editor[54] are developing support for BioPAX-based data exchange. In addition, tools for the storage and querying of Resource Description Framework (http://www.w3.org/RDF/) data sets, generated within the Semantic Web community, can be used to effectively process BioPAX data.

What is not covered?

The BioPAX language uses a discrete representation of biological pathways. Dynamic and quantitative aspects of biological processes, including temporal aspects of feedback loops and calcium waves, are not supported. However, BioPAX addresses this need by coordinating work (as described below) with the SBML and CellML mathematical modeling language communities[55,56] and a growing software tool set supporting biological process simulation[57]. Detailed information about experimental evidence supporting elements of a pathway map is useful for evaluating the quality of pathway data. This information is only included in BioPAX for molecular interactions, because that was already defined by the PSI-MI language[58] and it was reused The BioPAX work group makes use of PSI-MI–controlled vocabularies and other concepts and works with the PSI-MI work group to build these vocabularies in areas of shared interest, such as genetic interactions. Although BioPAX does not aim to standardize how pathways are visualized, work is coordinated with the

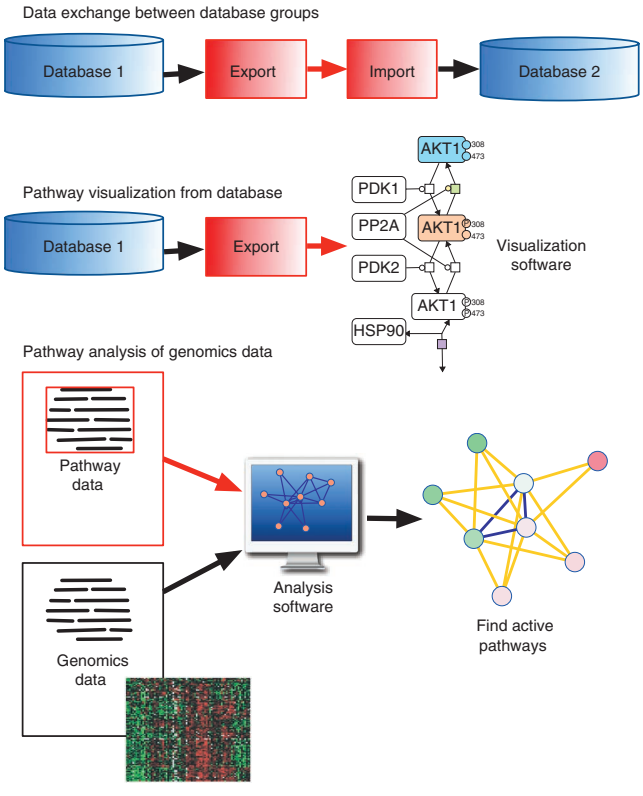

Figure 5 Example uses of pathway information in BioPAX format. Red-colored boxes or lines indicate the use of BioPAX.

Systems Biology Graphical Notation (SBGN; http://sbgn.org/) community, via members of both communities who attend BioPAX and SBGN meetings, to ensure that SBGN can be used to visualize BioPAX pathways. Currently, most BioPAX concepts can be visualized using SBGN process description and SBGN activity flow diagrams and a mapping of BioPAX to SBGN entity relationship diagrams is under development. BioPAX development is coordinated with the above standardization efforts through regular communication between workgroups to ensure complementarity and compatibility. For instance, controlled vocabularies developed by PSI-MI and BioPAX can be used to annotate SBML and CellML models (**Fig. 6**). BioPAX aims to be compatible with these and other efforts, so that pathway data can be transformed between alternative representations when needed. PSI-MI to BioPAX and SBML to BioPAX converters are available (**Supplementary Table 4**).

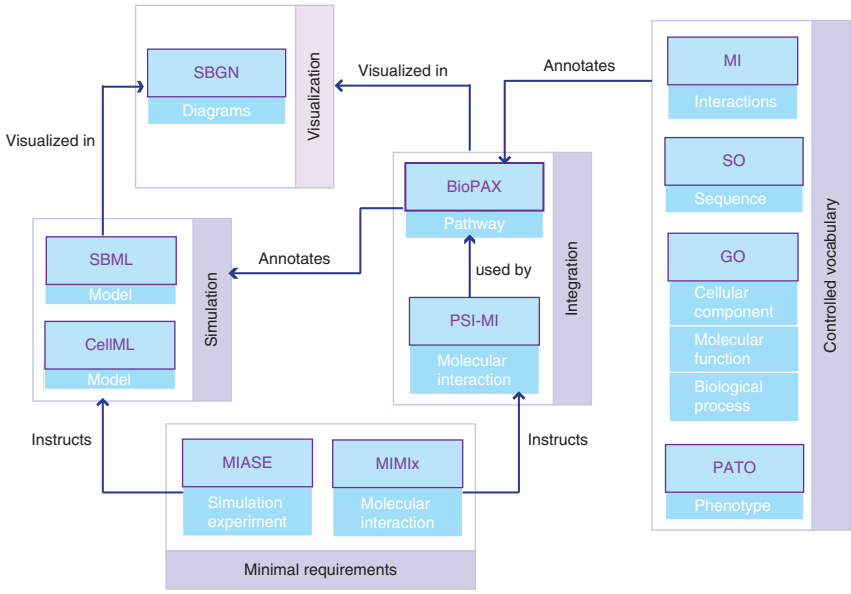

Figure 6 The relationship among popular standard formats for pathway information. BioPAX and PSI-MI are designed for data exchange to and from databases and pathway and network data integration. SBML and CellML are designed to support mathematical simulations of biological systems and SBGN represents pathway diagrams.

How does the BioPAX community work?

Whereas BioPAX facilitates communication of current knowledge, it is challenging for all knowledge-representation efforts to anticipate new forms of information. As new types of pathway data and new knowledge representation languages and tools become available, the BioPAX language must evolve through the efforts of a community of scientists that includes biologists and computer scientists.

BioPAX is developed through community consensus among data providers, tool developers and pathway data users. More than 15 BioPAX workshops have been held since November 2002, attended by a diverse set of participants. Incremental versions, also called levels, of the BioPAX language were progressively developed at these workshops to focus the group's efforts on attainable intermediate goals. Broader input came from mailing lists and a community wiki. Community members participated in developing functionality they were interested in, which was integrated into specific levels (**Supplementary Table 5**). Level 1 supports metabolic pathways. Level 2 adds support for molecular interactions and post-translational protein modifications by integrating data structures from the PSI-MI format. Level 3 adds support for signaling pathways, molecular state, gene regulation and genetic interactions (**Supplementary Table 3**). It is anticipated that newer BioPAX levels replace older ones, so use of the most recent BioPAX level 3 is currently recommended. To ease the burden on users and developers, BioPAX aims to be backwards compatible where practical. Level 2 is backwards compatible with level 1; however, level 3 involved a major redesign that necessitated breaking backwards compatibility. This said, many core classes have remained the same in levels 1, 2 and 3, and software is provided for updating older BioPAX pathways to level 3 (via Paxtools). All BioPAX material (**Table 1**) is made freely available under open source licenses through a central website (http://www.biopax.org/) to encourage broad adoption. The database and tool support (**Supplementary Table 4**) of a common language aids the creation, analysis, visualization and interpretation of integrated pathway maps.

In addition to the creation of a shared language for data and software, the process of achieving community consensus spurs innovation in the field of pathway informatics. Community discussion helps resolve technical knowledge representation issues faced by many data providers and users and facilitates the convergence to common terminology and representation. Solutions are discovered in independent research groups and incorporated in new data models and community best practices, which then enable identification of new issues. Thus, community workshops support a positive feedback cycle of knowledge sharing that has led to an accepted BioPAX language and development of better software and databases. We expect this to continue and to support new scientific uses of pathway information, motivated by end-user access to valuable integrated pathway information and efficiency gain for database and software development groups. This will especially benefit new pathway databases and software tools that adopt standard representation and software components from the start.

Future community goals

The BioPAX shared language is a starting point on the path to developing complete maps of cellular processes. Additional near and long-term goals remain to be realized to enable effective integration and use of biological pathway information, as described below.

Data collection. Data must be collected and translated to a standard format for them to be integrated. This process is underway, as the descriptions of millions of interactions in thousands of pathways across many organisms from multiple databases are now available in BioPAX format. However, vast amounts of pathway data remain difficult to access in the literature and in databases that don't yet support standard formats. Increasing use of standards requires promoting and supporting data curation teams and automating more of the data collection process using software. Easy-to-use tools for tasks like pathway editing must also be developed so that biologists can share their data in BioPAX format without substantial resource investment. Ideally, appropriate software would allow authors to enter data directly in standard formats during the publication process, to facilitate annotation and normalization by curators before incorporation into databases for use by researchers[53].

Validation and best practice development. To aid data collection, major data providers and others must develop community best practice guidelines and rules to help diverse groups use BioPAX consistently when multiple ways of encoding the same information exist. This will enable data providers to benefit from automatic syntactic and semantic validation of their data so they can ensure they are sharing data using standard representation and best practices[59,60]. Data collection and automatic validation will facilitate convergence to generally accepted biological process models.

Semantic integration. Several models of the same biological process may usefully co-exist. Ideally, different models could be compared for analysis and hypothesis formulation. Even so, comparison is difficult because the same concept can be represented in several ways owing to use of multiple levels of abstraction (such as the hRas protein versus the Ras protein family), use of different controlled vocabularies, data incompleteness or errors. Future research needs to develop semantic integration solutions that recognize and aid resolution of conflicts.

Visualization. Pathway diagrams are highly useful for communicating pathway information, but it is challenging to automatically construct these diagrams in a biologically intuitive way from pathway data stored in BioPAX. The SBGN pathway diagram standardization effort provides a starting point toward achieving this goal (**Fig. 3**). Intuitive and automatically drawn biological network visualizations may one day replace printed biology textbooks as the primary resource for knowledge about cellular processes.

Language evolution. As uses of pathway information and technology evolve, so must the BioPAX language. For instance, future BioPAX levels should capture cell-cell interactions, be better at describing pathways where sub-processes are not known or need not be represented, more closely integrate third-party controlled vocabularies and ontologies to ease their use and better encode semantics for easier data validation and reasoning.

Many groups within the BioPAX community, including most pathway data providers and tool developers, are working to achieve the above goals. For instance, Pathway Commons (http://www.pathwaycommons.org/) aims to be a convenient single point of access for all publicly accessible pathway information and the WikiPathways project (http://www.wikipathways.org/) seeks to enable pathway curation by individuals[53]. Also, the semantic web community is developing a set of technologies that promise to ease the integration of information dispersed on the World Wide Web[61]. These technologies will aid pathway data integration because BioPAX is compatible with them through use of the W3C standard Web Ontology Language, OWL. All of the above research and development activities support the vision of data providers sharing computable maps of biological processes in a standard format for convenient use by a community of pathway researchers.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

Funded by the US Department of Energy workshop grant DE-FG02-04ER63931, the caBIG program, the US National Institute of General Medical Sciences workshop grant 1R13GM076939, grant P41HG004118 from the US National Human Genome Research Institute and Genome Canada through the Ontario Genomics Institute (2007-OGI-TD-05) and US National Institutes of Health grant R01GM071962-07. Thanks to many people who contributed to discussions on BioPAX mailing lists, at conferences and at BioPAX workshops, especially A. Ruttenberg and J. Rees.

AUTHOR CONTRIBUTIONS

All authors helped develop the BioPAX language, ontology, documentation and examples by participating in workshops or on mailing lists and/or provided data in BioPAX format and/or wrote software that supports BioPAX. See **Supplementary Table 5** for a full list of author contributions.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788 (2003).
2. Nicholson, D.E. The evolution of the IUBMB-Nicholson maps. *IUBMB Life* **50**, 341–344 (2000).
3. Demir, E. *et al.* PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* **18**, 996–1003 (2002).
4. Krull, M. *et al.* TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.* **34**, D546–D551 (2006).
5. Fukuda, K. & Takagi, T. Knowledge representation of signal transduction pathways. *Bioinformatics* **17**, 829–837 (2001).
6. Davidson, E.H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
7. Kohn, K.W. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* **10**, 2703–2734 (1999).
8. Matthews, L. *et al.* Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37**, D619–D622 (2009).
9. Schaefer, C.F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–D679 (2009).
10. Bader, G.D. & Hogue, C.W. BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16**, 465–477 (2000).
11. Kitano, H. A graphical notation for biochemical networks. *BIOSILICO* **1**, 169–176 (2003).
12. Gama-Castro, S. *et al.* RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.* **36**, D120–D124 (2008).
13. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33**, D284–D288 (2005).
14. Keseler, I.M. *et al.* EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.* **37**, D464–D470 (2009).
15. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **38**, D473–D479 (2010).
16. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32** Database issue, D277–280 (2004).
17. Bader, G.D., Cary, M.P. & Sander, C. Pathguide: a pathway resource list. *Nucleic Acids Res.* **34**, D504–D506 (2006).
18. Huang, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
19. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007).
20. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
21. Karp, P.D. *et al.* Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.* **11**, 40–79 (2010).
22. Hu, Z. *et al.* VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res.* **35**, W625–W632 (2007).
23. Hoffmann, R. *et al.* Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE* **2005**, pe21 (2005).
24. Racunas, S.A., Shah, N.H., Albert, I. & Fedoroff, N.V. HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics* **20** Suppl 1, i257–i264 (2004).
25. Cary, M.P., Bader, G.D. & Sander, C. Pathway information for systems biology. *FEBS Lett.* **579**, 1815–1820 (2005).
26. Vivanco, I. & Sawyers, C.L. The phosphatidylinositol 3-Kinase AKT pathway in human cancer. *Nat. Rev. Cancer* **2**, 489–501 (2002).
27. Koh, G., Teong, H.F., Clement, M.V., Hsu, D. & Thiagarajan, P.S. A decompositional approach to parameter estimation in pathway modeling: a case study of the Akt and MAPK pathways and their crosstalk. *Bioinformatics* **22**, e271–e280 (2006).
28. Karp, P.D. An ontology for biological function based on molecular interactions. *Bioinformatics* **16**, 269–285 (2000).
29. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33** Database issue, D428–D432 (2005).
30. Mi, H., Guo, N., Kejariwal, A. & Thomas, P.D. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res.* **35**, D247–D252 (2007).
31. Demir, E. *et al.* An ontology for collaborative construction and analysis of cellular pathways. *Bioinformatics* **20**, 349–356 (2004).

32. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
33. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
34. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35**, D572–D574 (2007).
35. Kerrien, S. *et al.* IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.* **35**, D561–D565 (2007).
36. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539 (2006).
37. Matys, V. *et al.* TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
38. Kerrien, S. *et al.* Broadening the horizon—level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* **5**, 44 (2007).
39. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
40. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
41. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44 (2005).
42. Yamamoto, S., Asanuma, T., Takagi, T. & Fukuda, K.I. The molecule role ontology: an ontology for annotation of signal transduction pathway molecules in the scientific literature. *Comp. Funct. Genomics* **5**, 528–536 (2004).
43. Cerami, E.G., Bader, G.D., Gross, B.E. & Sander, C. cPath: open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics* **7**, 497 (2006).
44. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
45. Efroni, S., Carmel, L., Schaefer, C.G. & Buetow, K.H. Superposition of transcriptional behaviors determines gene state. *PLoS ONE* **3**, e2901 (2008).
46. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18** Suppl 1, S233–S240 (2002).
47. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
48. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.* **11**, R53 (2010).
49. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
50. Isserlin, R. *et al.* Pathway analysis of dilated cardiomyopathy using global proteomic profiling and enrichment maps. *Proteomics* **10**, 1316–1327 (2010).
51. Moraru, I.I. *et al.* Virtual Cell modelling and simulation software environment. *IET Syst. Biol.* **2**, 352–362 (2008).
52. Hlavacek, W.S. *et al.* Rules for modeling signal-transduction systems. *Sci. STKE* **2006**, re6 (2006).
53. Pico, A.R. *et al.* WikiPathways: pathway editing for the people. *PLoS Biol.* **6**, e184 (2008).
54. Kitano, H., Funahashi, A., Matsuoka, Y. & Oda, K. Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.* **23**, 961–966 (2005).
55. Lloyd, C.M., Halstead, M.D. & Nielsen, P.F. CellML: its future, present and past. *Prog. Biophys. Mol. Biol.* **85**, 433–450 (2004).
56. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
57. Sauro, H.M. *et al.* Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS* **7**, 355–372 (2003).
58. Hermjakob, H. *et al.* The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.* **22**, 177–183 (2004).
59. Racunas, S.A., Shah, N.H. & Fedoroff, N.V. A case study in pathway knowledgebase verification. *BMC Bioinformatics* **7**, 196 (2006).
60. Laibe, C. & Le Novere, N. MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Syst. Biol.* **1**, 58 (2007).
61. Berners-Lee, T. & Hendler, J. Publishing on the semantic web. *Nature* **410**, 1023–1024 (2001).
62. Le Novere, N. *et al.* The Systems Biology Graphical Notation. *Nat. Biotechnol.* **27**, 735–741 (2009).
63. Knublauch, H., Fergerson, R.W., Noy, N.F. & Musen, M.A. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. in *The Semantic Web–ISWC 2004: Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004: Proceedings* (eds. McIlraith, S.A., Dimitris Plexousakis, D. & van Harmelen, F.) 229—243 (Springer, 2004).
64. Sowa, J.F. *Knowledge Representation: Logical, Philosophical, and Computational Foundations* (Brooks/Cole, 2000).
65. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35**, D5–D12 (2007).

[1]Computational Biology, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [2]Center for Bioinformatics and Computer Engineering Department, Bilkent University, Ankara, Turkey. [3]SRI International, Menlo Park, California, USA. [4]Institute for Bioinformatics Research and Development, Japan Science and Technology Agency, Tokyo, Japan. [5]Université libre de Bruxelles, Bruxelles, Belgium. [6]European Bioinformatics Institute, Hinxton, Cambridge, UK. [7]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [8]NYU School of Medicine, New York, New York, USA. [9]National Cancer Institute, Center for Biomedical Informatics and Information Technology, Rockville, Maryland, USA. [10]Predictive Medicine, Belmont, Massachusetts, USA. [11]BIOBASE Corporation, Beverly, Massachusetts, USA. [12]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico. [13]Biomolecular Systems Laboratory, Boston University, Boston, Massachusetts, USA. [14]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [15]McKusick-Nathans Institute of Genetic Medicine and the Departments of Biological Chemistry, Pathology and Oncology, Johns Hopkins University, Baltimore, Maryland, USA. [16]Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico. [17]Artificial Intelligence Center, SRI International, Menlo Park California, USA. [18]No affiliation declared. [19]Donnelly Center for Cellular and Biomolecular Research, Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada. [20]Faculté de Médecine, Université Rennes 1, Rennes, France. [21]Rothamsted Research, Harpenden, UK. [22]Cell Signaling Technology, Inc., Danvers, Massachusetts, USA. [23]Broad Institute, Cambridge, Massachusetts, USA. [24]Center for Food Safety and Applied Nutrition, US Food and Drug Administration, Laurel, Maryland, USA. [25]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA. [26]Neurobiology, Neurodegeneration and Repair Laboratory, National Eye Institute, National Institutes of Health, Bethesda, Maryland, USA. [27]Department of Behavioral Neuroscience. Oregon Health & Science University, Portland, Oregon, USA. [28]US Environmental Protection Agency Durham, North Carolina, USA. [29]Mathematics & Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA. [30]University of Connecticut Health Center, Farmington, Connecticut, USA. [31]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA. [32]Lexikos Corporation, Boston, Massachusetts, USA. [33]Biotechnology Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA. [34]Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda Maryland, USA. [35]Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK. [36]Clinical Semantics Group, Lexington, Massachusetts, USA. [37]Center for Cell Analysis and Modeling, University of Connecticut Health Center, Storrs, Connecticut, USA. [38]Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland. [39]Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria. [40]Department of Bioinformatics, Maastricht University, Maastricht, The Netherlands. [41]University of Auckland, Auckland, New Zealand. [42]Syngenta Biotech Inc., Research Triangle Park, North Carolina, USA. [43]Department of Genetics, Stanford University, Stanford, California, USA. [44]Yale Center for Medical Informatics, Yale University, New Haven, Connecticut, USA. [45]Loyola Marymount University, Los Angeles, California, USA. [46]Physiomics PLC, Magdalen Centre, Oxford Science Park, Oxford, UK. [47]St. John's University, Jamaica, New York, USA. [48]Columbia University, New York, New York, USA. [49]SRA International, Fairfax, Virginia, USA. [50]Novartis Knowledge Center, Cambridge, Massachusetts, USA. [51]University of Ottawa, Ottawa, Ontario, Canada. [52]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [53]Vertex Pharmaceuticals, Cambridge, Massachusetts, USA. [54]Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, Wisconsin, USA. [55]Gladstone Institute of Cardiovascular Disease, San Francisco, California, USA. [56]Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York, USA. [57]Centre for Biomedical Informatics, School of Medicine, Stanford University, Stanford, California, USA. [58]Computational Sciences, Informatics, Millennium Pharmaceuticals Inc., Cambridge, Massachusetts, USA. [59]Center for Biomedical Informatics and Information Technology, National Cancer Institute, Bethesda, Maryland, USA. [60]Institute for Genomics and Systems Biology, The University of Chicago and Argonne National Laboratory, Chicago, Illinois, USA. [61]Total Gas & Power, Paris, France. [62]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan. [63]Biological Network Modeling Center, California Institute of Technology, Pasadena, California, USA. [64]Department of Bioinformatics, Göttingen, Germany. Correspondence should be addressed to G.D.B. (biopax-paper@biopax.org).