BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 15 2011, pages 2127–2133
doi:10.1093/bioinformatics/btr339

Systems biology

Advance Access publication June 7, 2011

The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways

Luca Beltrame[1,†], Enrica Calura[1,2,†], Razvan R. Popovici[3,†], Lisa Rizzetto[1], Damariz Rivero Guedez[1], Michele Donato[4], Chiara Romualdi[2], Sorin Draghici[4,*] and Duccio Cavalieri[1,*]

[1]Department of Pharmacology 'M.Aiazzi Mancini', University of Firenze, 50139 Firenze, [2]Department of Biology, University of Padua, 35121 Padova, Italy, [3]Miravtech Corporation, Troy 48083, MI and [4]Department of Computer Science, Wayne State University, Detroit 48202, MI, USA

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Many models and analysis of signaling pathways have been proposed. However, neither of them takes into account that a biological pathway is not a fixed system, but instead it depends on the organism, tissue and cell type as well as on physiological, pathological and experimental conditions.

Results: The Biological Connection Markup Language (BCML) is a format to describe, annotate and visualize pathways. BCML is able to store multiple information, permitting a selective view of the pathway as it exists and/or behave in specific organisms, tissues and cells. Furthermore, BCML can be automatically converted into data formats suitable for analysis and into a fully SBGN-compliant graphical representation, making it an important tool that can be used by both computational biologists and 'wet lab' scientists.

Availability and implementation: The XML schema and the BCML software suite are freely available under the LGPL for download at http://bcml.dc-atlas.net. They are implemented in Java and supported on MS Windows, Linux and OS X.

Contact: duccio.cavalieri@unifi.it; sorin@wayne.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 22, 2010; revised on March 16, 2011; accepted on May 27, 2011

1 INTRODUCTION

In a famous commentary regarding systems biology (Lazebnik, 2004), Yury Lazebnik, using the analogy between biological pathways and electronic circuits, proposed the use of standard procedures through which even a biologist—without any specific knowledge—could fix a radio. One of the most challenging goals of modern biology is to decipher and describe the complexity of cell systems, and what Lazebnik pointed out is that without the integration of knowledge coming from different fields of science, the efforts of reverse engineering the cell are destined to fail.

Recently, research on system biology has been characterized by an increasing number of efforts to define common languages for sharing information in multidisciplinary areas (Abbott, 1999) with the aim to develop tools to describe accurate models, run effective simulations, visualize, analyze and integrate high-throughput data. Networks defined on studies about the interactions occurring within cell macromolecules are key elements for this research.

Existing biological networks can be classified into four categories, depending on the nature of their nodes and their interactions: metabolic pathways, molecular interactions, gene regulatory networks and signaling pathways (Li and Davidson, 2009; Wang *et al.*, 2007). There are many public resources which store and share representations of these networks including KEGG (Kanehisa and Goto, 2000), Reactome (Vastrik *et al.*, 2007), Biocarta (www.biocarta.org), Pathway Commons (www.pathwaycommons.org) and Wikipathways (Pico *et al.*, 2008), yet currently there is no gold standard on how biological pathways should be represented. This shortcoming affects particularly signaling pathways: without solid, consistent and unambiguous representations, hypotheses and analyses based on them are affected by an inability to do any proper computational analysis in the worst case, and a loss of power, in the best case. Furthermore, a proper representation of a pathway is important to enable efficient knowledge management and integration of data coming from multiple sources. Recent efforts on the representation of pathways have followed two main trends: a proper graphical representation and a machine-readable format.

On the basis of the existence and the use of graphical and machine-readable formats, pathway representations can be classified into three categories (Pan *et al.*, 2003): static, providing a non-modifiable graphical representation; semi-dynamic, representing information not only as a graphical map, but also using a corresponding machine-readable format, which is not, however, strongly interconnected with the graph; dynamic, where the graphical representation format depends directly on the underlying data model, and thus any modification in the latter can be immediately translated to the former. At the time of writing, all pathway representations stored in public databases are either static or semi-dynamic.

The most recent example of a pure graphical representation is the System Biology Graphical Notation (SBGN)

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

(Le Novere *et al.*, 2009). SBGN splits the representation of a biological network into three different levels (the process definition, the entity relationship and the activity flow language). The three representations are constructed in order to capture different aspects of the biological systems, defining a set of glyphs and constraints to reduce ambiguity and improve interpretation. The resulting representations are highly informative, and SBGN quickly achieved a broad consensus in the scientific community. However, despite ongoing efforts (Czauderna *et al.*, 2010), an SBGN-dedicated pathway repository does not exist yet, and the conversion from the existing pathway representations to SBGN format is still difficult, due to the higher level of specifications and the deeper knowledge required.

Machine-readable formats, on the other hand, aim at creating a representation of the pathway that can be read and interpreted by computer programs and used to perform analyses or predictions. Many formats have been proposed through the years, such as the Systems Biology Markup Language (SBML) (Hucka *et al.*, 2003) and the Biological Pathways eXchange (BioPAX) (Luciano, 2005). Although these initiatives are successful results of the joint efforts of a wider community, they are still incomplete and no one has truly been adopted as a community-wide standard. This is mostly due to the fact that the underlying biological problem and its associated complexity is hard to model into a data format, especially with regards to signaling cascades. Signaling cascades are coherent patterns of expression that happen through a pathway in a specific condition.

The complexity of the modeling of signal transduction (signaling) in pathways is due to many factors. Signaling pathways are formal descriptions of the signaling processes by which a cell converts certain signals into others, involving interconnected, finely regulated structures that may present a high level of redundancy. Also, these structures are integrated in a larger system and are affected by environmental-dependent changes. A further source of complexity lies in the fact that different cell types express different genes (or the same entities may be present in different cellular compartments) and thus produce different proteins, leading to significant alterations in the design of a specific signaling pathway.

This complexity makes precise modeling extremely difficult. As a matter of fact, the current pathway databases represent pathways regardless of the cell type and tissue they occur in. An imprecise model affects the capabilities of the analyses carried out with it (such as the statistical power to detect the pathways that are significantly impacted in a given condition), since it will not make use of the complete biological information available. With the intention to be more informative, pathway analysis research in recent years moved from analysis of gene lists to more complex algorithms able to exploit the topology of networks. The extraction of the topological information from a biological pathway and their interpretation to obtain a network is not a trivial task and are still extremely dependent on the level of detailed information provided by the data format (Alves *et al.*, 2006; Draghici *et al.*, 2007; Kashtan *et al.*, 2004; Massa *et al.*, 2010).

Here, we present the Biological Connection Markup Language (BCML), a bioinformatic framework that allows to design and to manage the representation of signaling pathways in a visual and machine-readable format that respects the SBGN specifications. This format allows an unambiguous and fully dynamic representation of signaling pathways that is useful for both the biologists and bioinformaticians. In the framework, any type of biological information (either gene identifiers, publications or other information) can be added to the elements of a BCML pathway to provide additional information on their presence in the pathway and their relationships. Lastly, since BCML is a machine-readable format, the conversion into other data formats is straightforward, with the possibility of contextual selection of elements.

2 METHODS

2.1 XML schema

An XML schema describing the complete SBGN Process Description version 1.1 was implemented according to the SBGN documentation (Supplementary Material). An additional schema, implementing support for findings, was written as an extension of the main SBGN schema. A set of utilities to validate, filter and extract entities from the schema was also developed in the Java programming language (Supplementary Material).

2.2 Dataset preprocessing

A set of dendritic cell samples unstimulated or stimulated with poly(I:C) and hybridized on Affymetrix microarrays was retrieved from Array Express (accession code E-MEXP-1230). Data were then pre-processed and normalized using the Robust Multichip Average (Irizarry *et al.*, 2003) while at the same time updating the chip definitions with up-to-date versions linked to Entrez Gene IDs (Dai *et al.*, 2005). Preprocessing was carried out using RMAExpress, version 1.0.

2.3 Differential expression

Differentially expressed genes were calculated by running the Rank Product algorithm (Breitling *et al.*, 2004) on the dataset. Each donor was considered a separate origin. To ensure statistical significance, the maximum cut-off for the percentage of false positives (pfp) was set at 0.05. The computation was performed with the Bioconductor package 'RankProd' using the R programming language.

2.4 Pathway analysis

The definition of the Toll-like receptor 3 (TLR3) pathway as described by the DC-ATLAS initiative (Cavalieri *et al.*, 2010) was used to create a BCML implementation of the pathway. The existing definition was expanded to include entities belonging to additional cell types and tissues. The reference definition of the TLR3 pathway was retrieved from Reactome (http://www.reactome.org) as a gene list.

The BCML file was then converted to gene list and to a format usable by SPIA, the R implementation of impact analysis (Tarca *et al.*, 2009).

The hypergeometric test was carried out on the gene lists using R. SPIA was performed with a modified package (to support custom pathways), using the differentially expressed genes obtained from previous analyses.

3 IMPLEMENTATION

3.1 Data format

In order to provide a human- and machine-readable format that is able to properly represent and analyze pathways, and with the feature of being compliant with the SBGN specifications, we developed a data format based on the SBGN Process Definition (PD) 1.1 specification (Supplementary Material).

BCML was defined using an XML Schema (http://www.w3.org/XML/Schema; Supplementary Material), supporting the complete SBGN PD definition, including entities, interactions, rules and

restraints. The various SBGN elements are defined as XML tags, with additional properties stored in tag attributes.

In addition to a full implementation of the SBGN specification, BCML provides a series of optional features (defined as extensions of the main schema). First of all, BCML can include additional information on the entities that compose the network: each entity be described by a series of database identifiers, e.g. Entrez Gene or Uniprot accession numbers and each species can have its independent set of identifiers. Furthermore, condition-specific information, called 'Findings', can be associated to each entity or reaction. 'Findings' are collections of biological information that are relevant to that entity or reaction. The current specification includes support for organism, organism part (tissue), cell type, the specific biological environment in which the evidence was proven and the type of the experiment used to gather evidence. To reduce ambiguity and promote consistency among different 'Findings', the schema enforces a controlled vocabulary built from current medical ontologies.

The BCML schema also provides support to split pathways into subpathways called 'macro modules', representing independent units of a signaling pathway, following the proposal brought forward by the DC-ATLAS initiative (http://www.dc-atlas.net) as part of the DC-THERA network of excellence (http://www.dc-thera.org).

Lastly, even though it is mostly focused on the biological data and its analysis, BCML also contains support for a number of graphical hints, such as border, background and text colors of the elements (while the original SBGN specification is monochromatic). These hints are recognized and processed by the tools we developed that can read and parse BCML files.

All the additions to the SBGN specification are completely optional. The layer structure of the format allows the use of additional data types without affecting the SBGN compliance.

3.2 BCML software suite

To support the use of the format, we developed a series of tools to ensure the proper description, manipulation and visualization of pathways using BCML. Through the use of the software suite, a pathway described in BCML can be checked for consistency (validation), represented according to the SBGN specification (graphical representation), specific elements can be excluded or included according to the user's criteria (contextual selection), experimental measurements can be added to the representation and data analysis with different methods can be carried out.

3.2.1 Data validation We developed a validation tool to ensure that BCML files are well formed. First of all, the validator ensures that the BCML file is well formed according to the XML specification, and secondly the network is examined for consistency, using rules and constraints defined by SBGN. In case of improperly constructed files, the tool then reports to the user which elements are breaking the specification. Lastly, the tool ensures that identifiers for each element are unique and that isolated entities are not present.

3.2.2 Graphical representation Since pathways described in BCML are XML files, appropriate tools can convert the BCML representation of a pathway into a number of other formats suitable for direct graphical representation.

No additional requirements are needed to produce a graphical representation out of a BCML file, as the structure of the format already contains all the needed information, which is then used by the software suite to produce a SBGN-compliant graph. The current implementation converts BCML files into GraphML, a widely used format for graph representation. BCML files converted to GraphML can be opened by programs such as the yEd graph editor, where they can be exported to vector graphs or bitmap images (Supplementary Material).

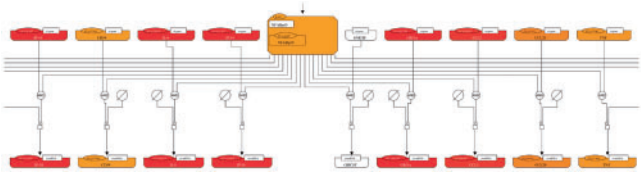

Fig. 1. Addition of experimental measurements to a BCML graphical representation. The diagram shows a section of the Toll-like receptor 3 pathway as described in BCML after addition of experimental measurements and graphical conversion. Red and orange indicate the degree of change (red highest and orange second highest, respectively). FC, fold change.

3.2.3 Contextual selection The tools present in the BCML software suite allow the on-the-fly customization of the pathway, taking advantage of the information stored in the format. For example, nodes and edges can be selected for a specific cell type or organism, allowing the construction of customized network maps to represent specific biological contexts. As an example, some elements of a specific pathway may be demonstrated in some cell types and not in others: BCML allows to create maps that include or exclude the elements that belong to a specific cell type.

The specification takes into account also elements and connections that may not be present given the selections, e.g. if in a specific cell type a complex may not form if one or more of its proteins are not present, marking them as 'affected by the selections', and providing a way to guide the analysis and the data interpretation and to point out gaps in current knowledge.

The results of the selection will also affect the graphical representation, as elements will be marked differently depending on their state (Supplementary Material).

3.2.4 Incorporation of experimental measurements Through the use of SBGN's StateVariable, the BCML format can incorporate any kind of experimental measurements that can be matched to the identifiers of an element. This permits, for example, to map high-throughput data coming from transcriptomics or proteomics experiments and to determine which elements of the pathway are affected in a given condition/tissue/organism. Measurements can be coupled with graphical hints so that when the pathway is converted to graphical representation, elements with experimental measurements will be colored accordingly (Fig. 1).

3.2.5 Support for analysis methods The presence of identifiers associated to the entities of a network described with BCML permits the transformation of the pathway in different data format suitable for data analysis. The tools provided with the suite permit the extraction of identifier (gene) lists from a BCML file, enabling their use with analysis methods such as Gene Set Enrichment Analysis (GSEA) and Fisher's exact test. Additionally, the format can be converted in a form amenable for impact analysis through the SPIA R package, enabling a topology-aware analysis of the network.

The conversion can take into account the contextual selection applied to the elements of the pathway, allowing analyses tailored for the user's experimental designs.

The tools are freely available and open source. A complete manual illustrating the use of the format and the tools is also included (Supplementary Material).

4 COMPARISON WITH ESTABLISHED FORMATS

4.1 Comparison with other formats

In order to assess the capabilities of BCML, we compared our format with the standards for the description of pathways that are most widely used by the scientific community: KGML, BioPAX and SBML (Table 1).

KGML is an XML-based format provided by the KEGG pathway database, defined by a specific Document Template Definition (DTD). KGML is the only pathway database currently offering pathway data in this format. Both BCML and KGML have a corresponding graphical visualization: in addition to what KGML offers, BCML has also support for compartment information and the additional features provided by SBGN. KGML is not SBGN compliant. The graphical representation of KGML is not derived by the format itself, but provided by KEGG directly as an image. This choice adds a manual process of translation from the format to the image, with potential discrepancy between the KGML entity and the graphical representation.

KGML and BCML support annotations for the entities of the pathway, in two different ways. KGML uses different files for different organism, while BCML incorporates these information in the format itself. When compared with KGML, BCML supports a wider number of annotations, such as cell type, organism and tissue.

KGML allows the coloring of the genes of the pathway through the supplied graphical representation: changes in interactions and/or in the pathway layout are not supported. BCML can instead use information not only to provide coloring of the entries (through the addition of experimental measurements), but also contextual selection of the elements basing on the annotation.

BioPAX is a format based on the Ontology Web Language (OWL), a derivative of XML based on subject–predicate–complement triplets, and is defined on multiple levels. Currently, it is used by many databases and resources, the most important being Reactome. Like BCML, it supports different types of pathways, including signaling pathways (starting from the recently released level 3 of the specification). It also supports annotation in a similar manner as BCML does. Both BCML and BioPAX supports contextual selection of the elements, in different ways: BioPAX, through the use of third-party tools, uses SPARQL to create a tabular output of the pathway, while BCML generates a second BCML file which is a subset of the initial pathway. This subset is a perfectly valid BCML file, which keeps all the features of the format. Therefore, when using BCML to perform contextual selection, no information is lost, as the original pathway is preserved.

With regard to graphical representation, BioPAX and BCML have a different design. BioPAX's official representation is SBGN: however, at the time of writing a tool that permits a direct conversion from BioPAX to SBGN and vice versa does not exist, although the BioPAX specification defines guidelines on how to convert from one format to another. External tools such as Cytoscape, through additional plug-ins, can produce a graphical visualization, although not SBGN compliant: for example, since Cytoscape models the pathway structure on a mathematical graph some structures required by SBGN, such as complexes, cannot be represented. On the other hand, as BCML is modeled on the SBGN specification itself, it can be automatically and directly transformed into a fully compliant graphical representation.

SBML is an XML-based purely computational format. It is used mainly to model biochemical reactions, due to its ability to annotate quantitatively the elements, but it also can be employed to describe signaling pathways. It is widely used in the scientific community and many different tools support this format. SBML, like BCML, supports annotations that can be added to elements using a controlled vocabulary. SBML is designed for the stoichiometry aspects of the reactions, while BCML is designated for the signaling aspects of the pathway. Additionally, SBML does not store the pathway layout in a standard fashion: the different tools that manipulate SBML either generate automatic layouts or use additional storage or definitions to store these information. BCML uses instead the terminology and the constraints of SBGN. Due to the way SBML is designed and its focus, contextual selection is not possible, because the model does not store biological information such as cell type or organism.

Table 1. Comparison of KGML, SBML, BioPAX and BCML

	KGML	SBML	BioPAX	BCML
Format	XML	XML	OWL-XML	XML
Schema freely available	Yes	Yes	Yes	Yes
Support for signaling pathways	Yes	No	Yes (level 3)	Yes
Contextual selection	No[a]	No	Yes[b]	Yes
Custom annotations	No	Yes	Yes	Yes
Graphical representation	Yes[c]	No	No	Yes
Support for experimental data	No[d]	Yes	No	Yes
Information about the biological environment	No	Yes	Yes	Yes
Description of experimental evidence	No	No	Yes	Yes

SBGN, Systems Biology Graphical Notation; KGML, KEGG Markup Language; XML, eXtensible Markup Language; OWL, Ontology Web Language.
[a]One version of each pathway for each supported organism.
[b]Produces tabular output.
[c]Limited to genes and reactions.
[d]Only different species are supported.

4.2 The Toll-like receptor 3 pathway as an example of BCML implementation

To properly evaluate the effectiveness of BCML in representing pathways, we retrieved the definition of the Toll-like receptor 3 (TLR3) pathway, a receptor involved in the immune recognition of double-stranded RNA, from two different sources: a reference implementation represented by the TLR3 pathway as stored in the Reactome database and a BCML version, as described in the Methods section. The BCML format was also transformed in its corresponding graphical output (Supplementary Material).

Table 2. Fisher's exact test results on the TLR3 pathway for its representation in BCML and Reactome

Pathway name	DEGs in pathway	Array genes in pathway	Array genes not in pathway	Total DEGs	*P*-value
TLR3 (BCML) filtered	24	69	17719	822	1.99840144433e-15
TLR3 (BCML) total	25	87	17701	822	8.21565038223e-14
Reference TLR3	9	74	17717	822	0.00525571923109

DEGs, differentially expressed genes.

Table 3. Impact analysis results on the TLR3 pathway represented in BCML

Pathway name	*P*-value	State
TLR3 (BCML) filtered	1.006106e-18	Activated
TLR3 (BCML) total	1.778768e-17	Activated

To determine the feasibility of the format for analysis methods, we tested BCML using a publicly available dataset (E-MEXP-1230 from Array Express), containing gene expression measurements on human dendritic cell samples stimulated with poly(I:C), a synthetic double-stranded RNA homologue that is recognized by TLR3. An additional feature of the BCML implementation was the ability of mapping differentially expressed genes (DEGs) directly inside the pathway format (Supplementary Material).

As a subsequent step, we converted our implementation to a gene list and performed the Fisher's exact test using the previously calculated DEGs alongside the Reactome reference. The reference implementation from Reactome yielded, as expected, a significant *P*-value of 0.00525. The test on the BCML implementation resulted in an improvement of the result significant *P*-value ($P = 8.21565038223e\text{-}14$; Table 2). The difference between the two results was due to the different level of curation in DC-ATLAS and Reactome.

In order to demonstrate the importance of having dynamic pathways and more informative data formats, we also tested the effect of contextual selection on the analysis (Table 2). The BCML implementation of the TLR3 pathway was modified selecting only the elements and reactions with evidence in human dendritic cells like the experimental setup, and analyses were reperformed. Results showed the *P*-values more significant by an approximately one order of magnitude indicating that more specific pathways lead to more informative analyses.

Due to its highly informative format, BCML is able to support automated analysis also on topology-dependent algorithms. Thus, we analyzed the data using a method that takes into account the relationships and the connections among elements (Impact analysis). We obtained a significant *P*-value and information on the trend of the data (pathway activated; Table 3). Analysis on the filtered TLR3 pathway yielded lower *P*-values than the complete pathway, an indication that filtering increases sensitivity and specificity. The results were in agreement with the Fisher's test.

These findings indicate that BCML format is as flexible as the already established public formats but also offers additional functionality missing in the publicly available counterparts.

5 DISCUSSION

In this work, we have presented BCML, a new data format designed for the representation of process description specification of the SBGN data model for the representation of biological networks. Moreover, our format was designed to provide significant additional capabilities, dramatically increasing the amount of information that it can store and offering additional flexibility. These additional capabilities are useful for both biological interpretation and data analysis.

The aim that guided the development of BCML was to build a flexible and dynamic representation of biological pathways in an unambiguous way, while still being understandable by the biologist, being able to improve the available analysis and supporting features such as contextual selection, extended annotation and graphical visualization. We chose XML because it is highly flexible, easily parseable and simply transformable into other formats. Moreover, XML is an easy to learn language and a well-designed schema (e.g. using descriptive and self-explanatory tags) allows an easy to understand reading of the files produced with it, also for people with limited computer skills (Barillot and Achard, 2000). Also, specialized XML editors can create easily understandable data models composed of entities nested on multiple levels. Such available editors also facilitate XML content visualization.

The adoption the SBGN model ensures that BCML is able to represent the major part of the bulk of biological information in an unambiguous way, and at the same time, is compatible with the most recent graphical standard for biological pathways.

Before developing a new format, we examined KGML, BioPAX (Luciano, 2005) and SBML (Hucka *et al.*, 2003). Each of the formats provided features that we needed, but neither covered exactly our use case. SBML had a different focus, modeling quantitative and temporal aspects, than our intended goal and did not properly support a graphical representation. KGML was not detailed enough to capture all the required information, and BioPAX did not implement at the time of writing some of our requirements, for example generation of subpathways from contextual selection. Also, KGML and SBML did not offer SBGN-compliant representations, while BioPAX required an additional step of conversion.

One of the most important problems we faced when building this model was the fact that in general terms, most pathways available in public databases are 'generic'. Although most pathways are organism specific, they lack information on the precise cellular types or tissues in which the described phenomena take place. Even more seriously, many publicly available pathways combine in a single diagram elements that are specific only to certain tissues (Cavalieri *et al.*, 2010). Since the commonly used pathway databases do not currently store information about tissue localization, a life scientist unfamiliar with the minutiae of a specific pathway could easily and incorrectly infer that these genes would be involved in this pathway in all cells. In contrast, a tool relying on BCML can display (on request) only those elements of the pathways that are known to be applicable in the currently selected organism and tissue type,

and supported by the level of proof specified by the user. Thus, one single BCML pathway representation can capture all the information available about that pathway for all organisms, tissue types, type of evidence, etc., as well as experimentally measured values such as gene expression, protein abundance, etc.

We are strongly convinced that the most improvements in the analysis of pathway will come from a better use of the already existing knowledge. For this reason, we needed formats able not only to store additional information (findings) about a pathway, but also to be able to include or exclude elements based on such existing knowledge thus enabling the creation of 'customized' pathways, better suited to describe specific biological problems or to highlights gaps in our current knowledge.

BCML is our proposal to overcome this lack of flexibility in the current available data formats for pathways. The constraints set by BCML, which reflect the ones set up by SBGN, are also important in ensuring that a pathway will be designed in the correct form from the start. Moreover, in BCML we added the possibility to integrate experimental measurements as a way to improve interpretation of experimental results.

Lastly, we wanted to construct a format that could facilitate subsequent analyses, because the use of pathways to perform analysis, especially in the context of high-throughput data, is an expanding field (Cavalieri and De Filippo, 2005; Werner, 2008). BCML can, through very simple transformations, be used both for gene list-based approaches such as GSEA (Subramanian *et al.*, 2005) or the canonical Fisher's Exact Test (Draghici *et al.*, 2003; Grosu *et al.*, 2002), and as well as the more advanced, topology-aware methods such as Impact Analysis (Tarca *et al.*, 2009). Furthermore, the contextual selection can also be applied when transforming the data for analysis, thus permitting analyses 'tailored' to specific biological problems.

In order to ensure that our format had an advantage, according to our design, to the already existing formats, we compared two specific implementations of the Toll-like receptor 3, a receptor involved in the dendritic cell response to double-stranded RNA (Kawai and Akira, 2007; Meylan and Tschopp, 2006): one represented using our format, and as the other we used the one stored in the Reactome pathway database, which we used as established reference. We then used a publicly available dataset to test the reliability of the implementation applied to data analysis: we expected an activation of the TLR3 pathway, because the dataset contained dendritic cells stimulated with poly(I:C), a synthetic homolog of the dsRNA recognized by the receptor.

Our results showed that the BCML representation of the TLR3 pathway performs as well as the established reference, while offering important additional features, such as the possibility of incorporating experimental measurements, the possibility of using topology-aware analysis algorithms and the contextual selection of elements according to a specific biological context: when compared to the use of different pathways in the same species, such an approach is more powerful as it can be used to highlight subtle differences in signaling networks among different cell types or tissues.

With regards to classical analysis methods, both Reactome and the BCML implementations of TLR3 gave significant *P*-values for TLR3 pathway. The discrepancy in the results of the statistical test is not related to the format, but it is an expression of the different curation in the standard (Reactome) and DC-ATLAS (Cavalieri *et al.*, 2010), where the BCML implementation of TLR3 was taken from. This shows that our implementation of TLR3 is comparable to established standards when using an external, focused dataset. Additionally, BCML provides the possibility of using topology-aware analysis methods such as Impact Analysis, which are more precise as they take into account the order and the causal relationships among the various entities.

For both methods, using a subset of the TLR3 pathway produced by contextual selection, keeping into account the specific biological context, yielded *P*-values lower by one order of magnitude with respect to the generic implementation. This result is of particular importance because it clearly shows the need for pathway definitions that match as closely as possible the biological context that is being investigated, leading to more robust and precise results. As a matter of fact, pathways and cellular networks exhibit even greater differences between species (Mestas and Huges, 2005; Shen-Orr *et al.*, 2010), and even more importantly, the cell type is an even greater discriminating factor: for example, in the TLR3 pathway, the presence of $>50\%$ of its known genes has not been demonstrated in dendritic cells (Cavalieri *et al.*, 2010). Thus, when using the pathway definition for computational analysis, it is essential to be as close as possible to the experimental setup to prevent or notice inconsistencies that will ultimately affect the final interpretation of the results. Despite BCML's lack of refinement compared to the currently available alternatives, it provides additional functionality, and it highlights a possible solution to problems that are now arising when representing pathways. Thanks to the selection capabilities of BCML, it is possible to construct specific pathways for 'tailored' analyses. Such selection can be used both by the biologist to visualize the non-demonstrated interactions and to the bioinformatician who can adjust the analysis methods to take missing annotations into account.

BCML only covers Process Description at the moment, while a complete SBGN representation of the pathway should also includes the Entity Relationship and Activity Flow representation. We will work toward implementing these two specification in order to provide a more complete data model for all the SBGN graphical pathway descriptions. We will also work on developing tools to convert BCML to other formats such as BioPAX to increase interoperability.

One of the major strengths of this work is that the format was conceived in parallel with the implementation of flexible tools for the representation, manipulation and analysis of biological networks. We are hereby making available to the community, not only an abstract data model for a possible large adoption, but also the basic tools that allow the manipulation of pathways in this format.

Being SBGN compliant and machine readable, BCML provides a convenient and precise way to represent biological pathways, in a form useful to both the biologist and the bioinformatician. Its dynamic nature makes it an important tool for the dissection of complex, highly specific biological problems. Lastly, BCML containing deeper descriptions of biological knowledge turns out to be a format extremely suitable for advanced pathway analysis methods, as well as creation of knowledge-based online resources. We expect that our model will be a useful contribution to the pathway community, making possible the creation of more practical and more complete pathway representations that will be both more end-user friendly, as well as better suited for advanced computational analysis.

ACKNOWLEDGEMENTS

We are grateful to the SBGN consortium for the helpful discussions.

Funding: This work was supported by grants EU LSHB-CT-2004-512074 (DC-THERA) and 242220 (SYBARIS European Network of Excellence).

Conflict of Interest: none declared.

REFERENCES

Abbott,A. (1999) Alliance of US labs plans to build map of cell signalling pathways. *Nature*, **402**, 219–220.

Alves,R. *et al.* (2006) Tools for kinetic modeling of biochemical networks. *Nat. Biotechnol.,* **24**, 667–672.

Barillot,E. and Achard,F. (2000) XML: a lingua franca for science? *Trends Biotechnol.*, **18**, 331–333.

Breitling,R. *et al.* (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.

Cavalieri,D. and De Filippo,C. (2005) Bioinformatic methods for integrating whole-genome expression results into cellular networks. *Drug Discov. Today*, **10**, 727–734.

Cavalieri,D. *et al.* (2010) DC-ATLAS: a systems biology resource to dissect receptor specific signal transduction in dendritic cells. *Immunome Res.,* **6**, 10.

Czauderna,T. *et al.* (2010) Editing, Validating, and Translating of SBGN Maps. *Bioinformatics*, **26**, 2340–2341.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, 175.

Draghici,S. *et al.* (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Draghici,S. *et al.* (2007) A systems biology approach for pathway level analysis. *Genome Res.*, **17**, 1537–1545.

Grosu,P. *et al.* (2002) Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res.*, **12**, 1121–1126.

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.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kashtan,N. *et al.* (2004) Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs. *Bioinformatics,* **20**, 1746–1758.

Kawai,T. and Akira,S. (2007) Antiviral signaling through pattern recognition receptors. *J. Biochem.*, **141**, 137–145.

Lazebnik,Y. (2004) Can a biologist fix a radio? – Or, what I learned while studying apoptosis. *Cancer Cell*, 2002, **2**, 179–182. *Biochemistry (Mosc)*, **69**, 1403–1406.

Le Novere,N. *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, **27**, 735–741.

Li,E. and Davidson,E.H. (2009) Building developmental gene regulatory networks. *Birth Defects Res. C Embryo Today*, **87**,123–130.

Luciano,J.S. (2005) PAX of mind for pathway researchers. *Drug Discov. Today*, **10**, 937–942.

Meylan,E. and Tschopp,J. (2006) Toll-like receptors and RNA helicases: two parallel ways to trigger antiviral responses. *Mol. Cell*, **22**, 561–569.

Massa,M.S. *et al.* (2010) Gene set analysis exploiting the topology of a pathway. *BMC Syst. Biol.,* **4**, 121.

Mestas,J. and Hughes,C.C. (2004) Of mice and not men: differences between mouse and human immunology. *J. Immunol*., **172**, 2731–2738.

Mestas,J, *et al.* (2005) Endothelial cell co-stimulation through OX40 augments and prolongs T cell cytokine synthesis by stabilization of cytokine mRNA. *Int. Immunol.,* **17**, 737–47.

Pan,D. *et al.* (2003) PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. *BMC Bioinformatics*, **4**, 56.

Pico,A.R. *et al.* (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.

Shen-Orr,S.S. *et al.* (2010) Cell type-specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tarca,A.L. *et al.* (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.

Vastrik,I. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.

Wang,E. *et al.* (2007) Cancer systems biology: exploring cancer-associated genes on cellular networks. *Cell Mol. Life Sci.*, **64**, 1752–1762.

Werner,T. (2008) Bioinformatics applications for pathway analysis of microarray data. *Curr. Opin. Biotechnol.*, **19**, 50–54.