# BiNoM, a Cytoscape plugin for accessing and analyzing pathways using standard systems biology formats

Eric Bonnet[1,2,3] , Laurence Calzone[1,2,3] , Daniel Rovera[1,2,3] , Gautier Stoll[1,2,3] Emmanuel Barillot[1,2,3] and Andrei Zinovyev[1,2,3*]

[1]Institut Curie, 26 rue d'Ulm, Paris, F-75248 France
[2]INSERM, U900, Paris, F-75248 France
[3]Mines ParisTech, Fontainebleau, F-77300 France

Email: andrei.zinovyev@curie.fr;

*Corresponding author

## Abstract

Public repositories of biological pathways and networks have greatly expanded in recent years. Such databases contain many pathways that facilitate the analysis of high-throughput experimental work and the formulation of new biological hypotheses to be tested, a fundamental principle of the Systems Biology approach. However, large scale molecular maps are not always easy to mine and interpret. We have developed BiNoM (Biological Network Manager), a Cytoscape plugin which provides functions for the import-export of standard Systems Biology file formats (such as SBML and BioPAX), and a set of algorithms to analyze and reduce the complexity of large biological networks. BiNoM can be used to import and analyze files created with the CellDesigner software. BiNoM provides a set of functions allowing to import BioPAX files, but also to search and edit their content. As such, BiNoM is able to efficiently manage large BioPAX files such as whole pathway databases. BiNoM also implements a collection of powerful graph-based functions and algorithms such as path analysis, decomposition by name or cycle, subnetworks clustering and decomposition of a large network in modules. Here, we provide an in-depth overview of the BiNoM functions, and we also detail novel aspects such as the support of the BioPAX level 3 format and the implementation of a new algorithm for the quantification of pathways for influence networks. At last, we illustrate some of the BiNoM functions on a detailed biological case study of a network representing the G1/S transition of the cell cycle, a crucial cellular process disturbed in most human tumors.

## Background

Biological pathways and networks comprise sets of interactions, or functional relationships, occurring at the molecular level in living cells [1, 2]. A large body of knowledge on cellular biochemistry is organized in publicly available repositories such as the KEGG database [3], Reactome [4], MINT [5], or the Cancer Cell Map (http://cancer.cellmap.org/cellmap/). All these pathway and biological network databases facilitate a large spectrum of analyses, improving our understanding of cellular systems. For example, it is now a very common practice to cross the output of high-throughput experiments, such as mRNA or protein expression levels, with curated biological pathways in order to visualize changes, analyze their impact on a network and formulate new hypotheses about biological processes [6,7]. The development of those pathway repositories has also fueled the creation of standard representations and formats, to facilitate the exchange and representation of data, such as the Biological Pathway Exchange standard (BioPAX) [8], the Systems Biology Markup Language (SBML) [9] or the Systems Biology Graphical Notation (SBGN) [10]. The Pathguide website counts more than 300 web-accessible biological pathway and network databases [11], many of which are using the SBML and BioPAX standard formats. Ultimately, those integrated resources will facilitate computational models building, their exchange, re-usability and their experimental validation, a cycle that is the cornerstone of the Systems Biology approach [12–14].

As a consequence, there is a need for the precise and accurate construction of pathways and large scale molecular maps covering fundamental biological processes. Such maps are often constructed by manual curation of the literature or automated curation from pathway databases [15]. More and more, they are focused on the regulation of biological processes involved in diseases such as cancer, Alzheimer's disease or Crohn's disease, to name a few [16–19]. However, the scale of such maps, even when they are focusing on a particular process, is quite large, with hundreds of chemical species and interactions. The analysis and interpretation of such large scale maps is therefore not a straightforward task. Several computational tools have been developed to facilitate the visualization, curation and analysis of pathways [1]. For example, CellDesigner is a software package for the graphical editing of biological pathway diagrams [20], using a proprietary extension of SBML to store all the information contained in the graphs. There is obviously a need for user-friendly software tools that would allow the user to easily import data from various standard format sources, to perform structural analyses on these pathways and to manipulate networks, and to be able to export a network to a suitable format for further analysis (e.g. mathematical modeling). We have created BiNoM [21], a software plugin for the popular Cytoscape network vizualization and analysis tool [22]. BiNoM has several built-in functions for importing and exporting standard Systems Biology file formats

such as SBML and BioPAX. BiNoM also implements several functions based on graph operations for the structural analysis of biological networks. Those functions can be used to reduce the complexity and extract meaningful subnetworks from large scale molecular maps. Here, we provide a detailed view on the functions implemented in BiNoM that permit specific extraction of information from large scale molecular maps and improve their readability and usability. We also highlight novel functions that were implemented recently, such as the support of the latest BioPAX specification (BioPAX level 3) and an algorithmic approach for the quantification of pathways on influence networks (PIQuant). We illustrate the use of the principal BiNoM functions with a detailed analysis of a molecular netwok of the G1/S transition of the cell cycle, a central mechanism for tumor development and progression.

## Implementation

BiNoM is implemented in the Java programming language, as a plugin for the network visualization and analysis software package Cytoscape [22]. Although the primary use of BiNoM is through the Cytoscape software, the underlying logic of most of the BiNoM functions is completely decoupled from the Cytoscape objects, allowing developers to also use BiNoM as an independent java library [21]. The installation of BiNoM can be done through the Cytoscape plugin manager (menu "Plugins > Manage Plugins", section "Other"). Alternatively, the user can also download the plugin together with a manual and the source code from the BiNoM website http://binom.curie.fr/.

BiNoM is designed to handle different Systems Biology file formats and to provide useful functions for the analysis of biological networks. The core functions of BiNoM can be grouped in five different topics: Input/Output, Analysis, BioPAX utils & query, Module manager and Utilities.

### BiNoM Input/Output

BiNoM functions facilitate the import and export of standard Systems Biology file formats, but BiNoM plugin is not designed to be a universal converter. The goal is more to provide robust functions for a finite number of useful conversion and analysis scenarios, such as the examples showed below (non-exhaustive list):

- Interconversion of CellDesigner files to BioPAX, and from a BioPAX reaction network to SBML level 2.
- Import of a BioPAX file as a reaction network and/or a pathway structure and/or a protein interactions network, followed by the creation of a subnetwork saved as a new BioPAX file.

- Import of a BioPAX file, selection of a subnetwork of interest saved as a SBML file for the creation of a computational model using an appropriate software package.
- Import a large CellDesigner map and export only a subnetwork as a new CellDesigner file.

BiNoM allows the user to import and export SBML level 2 files, as well as CellDesigner 3.x and 4.x file formats [21]. The BioPAX community has recently made a major update of the BioPAX standard, producing a new specification known as BioPAX Level 3 (http://www.biopax.org/). This format supports metabolic pathways, signaling pathways (including states of molecules and generic molecules), gene regulatory networks, molecular interactions and genetic interactions. Due to major changes in the specification, the BioPAX level 3 is not backward compatible with the level 2 file format.

BioPAX is using the web ontology language specification (OWL, http://www.w3.org/2004/OWL/) to store data in XML-formatted files. In BiNoM, we use the Jastor and the Jena java libraries (http://jastor.sourceforge.net/, http://jena.sourceforge.net/) to automatically create java classes from the BioPAX specifications, allowing a convenient access to the different data types encoded in the BioPAX files. The guiding principle in BiNoM for the access to Systems Biology files is to provide control over the content without completely converting the file to the Cytoscape format. The content is therefore mapped to a labeled directed graph, representing the complete set of objects and their relationships. This graph, called the index, is highly connected and is not visualized explicitly. Instead, the user interacts with selected subgraphs extracted from the index. For example, BioPAX files are imported as three separate graphs, respectively the Reaction Network (RN), representing the biochemical reaction network, the Pathway Structure (PS), showing the hierarchical organisation of pathways, and the Interaction Map graph (IM). Several examples of simple BioPAX level 3 files imported through BiNoM, representing different types of interactions, are shown on figure 1. Figure 2 shows the hierarchical structure of the human Apoptosis pathway, extracted from Reactome database, and constructed by BiNoM.

When importing a file, BiNoM is calling a *naming service* function in order to create meaningful names for the various entities. More precisely, entity names are combined with other features such as modifications, compartment and complex components. The different features are indicated by special characters, such as "@" for the compartments, "|" for modifications and ":" to delimitate the different members of a complex. For example, the name Cdc25|Pho@cytoplasm represents the protein Cdc25 in a phosphorylated state, located in the cytoplasm, while the name Cdc13:Cdc2|Thr167_pho@cytoplasm indicates a protein complex located in the cytosplasm, composed of the protein Cdc13 and the protein Cdc2 phosphorylated at position

167 on a threonine residue.

### BiNoM Analysis

The central goal of the BiNoM plugin is to provide efficient methods and algorithms to reduce the inherent complexity of biological networks into manageable and meaningful subnetworks. This goal is achieved by a set of functions included as a built-in structural graph analysis library. Some of the functions take into account the semantics contained in the graph element names. The structural analysis functions implemented in BiNoM include the analysis of connected and strongly connected components, pruning the network, decomposition by involvement of a protein (material components) or by cyclic decomposition, path analysis and network clustering. We also introduce in this version of BiNoM a novel function to quantify the influence of a source node on a target node taking into account experimental data, called PIQuant.

#### *Decomposition by involvement of a protein or by cyclic decomposition*

BiNoM allows many methods to dissect a complex biological network into parts. A trivial approach to separate a network into subparts is to dissociate the unconnected subparts of the network. A more sophisticated one consists in decomposing the network into strongly connected components, using the algorithm of Tarjan [23]. It is also possible to prune the network into three different parts: the one with all the elements associated with the *input* part of the network (from which all paths lead to the central core), the second with all the elements associated with the *output* part (from which there are no paths leading back to the central core) and the last part with all the elements linked to the central core, cyclic part, composed from strongly connected components, possibly connected together. This type of approach corresponds to finding the bow-tie graph structure [24].

The decomposition in material components is using the node name semantics to isolate subnetworks in which each protein is involved, either as a simple chemical species or as part of a complex. As a result, major overlaps between the different subnetworks are to be expected, as many proteins are expected to be involved in different complexes. Figure 3 shows two examples of subnetworks obtained by material component decomposition applied to a cell cycle network model of the yeast species *S. pombe* [25]. This approach represents different parts of the life cycle of a given protein.

The cycle decomposition is splitting the network into relevant directed cycles [26], using a modifed version of the algorithm of Vismara and colleagues [27]. This procedure commonly shows the different mechanisms in which the protein is playing a role. Care must be taken when applying this approach, as the number of

cycles can be huge for large network structures. For example, it might be preferable to eliminate first the network hubs, which are by definition highly connected, and also group short cycles in larger subnetworks before applying the decomposition function. Figure 4 shows two cycles involving CDC25 after a cycle decomposition.

***Path analysis algorithms***

BiNoM analysis functions also include classical path analysis algorithms, such as finding the shortest paths, the suboptimal shortest paths or all non self-intersecting paths (table 1). The shortest path is calculated as the path having the minimal sum of weights of the edges composing the path (Dijkstra's algorithm) while the suboptimal path is constructed by removing all edges of all shortest paths one by one, and finding the new shortest path. All non self-intersecting paths are those paths not containing loops (self-intersections). The user should be careful when using this procedure, as the number of paths between nodes can be very large for big networks. In order to limit the number of paths found, BiNoM allows to specify the maximal length of the path to be found.

Obviously, the result of some decomposition functions will result in subnetworks that share some components, as it is for example often the case with the decomposition in material components. Therefore, BiNoM also includes a function to *cluster* networks, based on common components such as protein or protein complexes. To determine the size of the clusters, the user can specify a percentage of intersection (ranging from 0 to 100%) that will be used as a threshold to create the clusters.

***Pathway influence quantification algorithm***

In this version of BiNoM, we have introduced a novel approach to quantify the effect of experimental data onto one or more target nodes for a given network architecture, named the PIQuant score (Pathway Influence Quantification). The target node can be a gene or a phenotype of interest, representing a more complex biological function, such as cell proliferation or apoptosis. For example, in the case of a network with one source node and one target node, the PIQuant score value will quantify the influence of the source node on the target node, by taking into account experimental data values for the input nodes, the path length and the sign of the path for a set of paths defined between the source and the target node (see a summary of the input data types for the PIQuant score in table 2). A positive or negative PIQuant score value is a quantitive theoretical prediction of the over or underexpression of the target node. For instance, let us consider that we have experimental data for a given network corresponding to differential

gene expression values (e.g. disease/normal ratios). In that case, a positive or a negative PIQuant score for a given phenotype (output node) predicts quantitatively that the phenotype would be respectively enhanced or inhibited. Furthermore, the relative difference between two PIQuant score values also is also indicative of a relative quantitative difference. Thus, the PIQuant score can be used to compare the effects of two different experimental datasets on the same phenotype (i.e using the same network), or to compare the effects of two different network architectures on the same phenotype for one experimental dataset.

More formally, we define a source node as *annotated* when a signed real number is assigned to the node, representing an experimental data value (e.g. the expression ratio of a gene between a disease and a normal state, obtained from transcriptomics profiling). A path $k \in \{1, \ldots, q\}$ is defined as a the sequence of consecutive connected nodes between a source node and a target node (without repetition of any node or edge). We can extract a set of paths from source nodes to target nodes (indexed from 1 to $q$), by using various algorithms. In BiNoM, we propose three solutions to search for paths between the source and target nodes (shortest paths, suboptimal shortest paths and all non-intersecting paths, see previous paragraph). The activity $\alpha_k$ of the path $k$ is defined as the annotation of its source node. We define the sign $\sigma_k$ of the path $k$ as the product of the signs of every edge of the path and finally the length $\lambda_k$ of the path $k$ as the number of edges in the path. We hypothesize that the longer the path is, the lesser the global influence will be on the target node. This assumption has the advantage of being simple and does not require extra parameters other than the influence network to be calculated. We assume that we a set of $q$ paths that have been extracted from the network of interest, between a selection of input and output nodes defined by the user. The PIQuant score is then defined as:

$$PIQuant_{Score} = \sum_{k=1}^{q} \alpha_k \sigma_k \frac{1}{\lambda_k}$$

In the case of the network presented in figure 5a, let us consider Ac the source node and Ph the target node and consider only the two paths defined in the Figures 5b and 5c. Given that the node Ac is annotated by the value 2.0, that the first path has a length equal to 3, and that the second path has a length equal to 5, we can calculate the PIQuant score of the node Ac to the node Ph as:

$$PIQuant_{Score} = 2 \cdot 1 \cdot \frac{1}{3} + 2 \cdot (-1) \cdot \frac{1}{5} = 0.27$$

Sometimes a path has one or more intermediate nodes that are annotated nodes (i.e. for which we have experimental data values), in addition to the source node. In that case, the intermediate node is *consistent*

if the sign of its annotation is the same as the sign of the source node annotation multiplied by the sign of the path from the source node to the intermediate node. If signs are opposite, the node is *inconsistent*. A path is *consistent* if each annotated intermediate nodes is consistent. A path is *inconsistent* if at least one intermediate node is *inconsistent*. In other words, a path is inconsistent when the sign of the experimental data value for a node is opposite to the sign of the path.

For example, according to the network represented in figure 5c, there is an influence from Ac to Ph that goes through D . But if this path is functional, the annotation of node D should have the same sign as the annotation for node Ac, because Ac activates D indirectly (i.e. the path from Ac to D has a positive sign). In this case, the sign of D annotation (the experimental value) is opposite to the sign of Ac annotation, and therefore, the path from Ac to Ph that goes through D is inconsistent.

Practically, we offer the option to keep or not the inconsistent paths for the calculation of the PIQuant score, depending on how the user wants to analyze the calculations. An inconsistent path could indicate that the path is not complete, or could also indicate that the path is correct but not active under the precise conditions in which the experimental data was generated, corresponding to a different context.

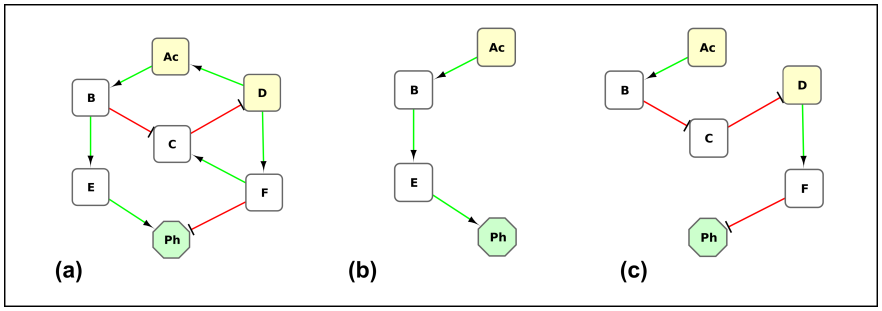
In the case mentionned above in figure 5c, if only consistent paths are kept, then the PIQuant score of Ac to Ph becomes:

$$PIQuant_{Score} = 2 \cdot 1 \cdot \frac{1}{3} = 0.67$$

This score is higher than the value previously obtained with all the paths, and can be interpreted in this case as a higher activation of the phenotype.

In more realistic situations, we would have multiple source nodes with different annotations, and also multiple target nodes representing phenotypes of particular nodes of interest. We have implemented in BiNoM a set of functions that allow users to select source nodes, select target nodes, and choose among three different options for searching paths (shortest paths, optimal and suboptimal shortest paths, all non-intersecting paths). The software is then calculating PIQuant scores for every target node specified, taking into account every possible path found by the algorithm. An interactive window is detailing the PIQuant score results, both globally and for every path from the source nodes to the output nodes. It is also possible to get a full text report detailing all the calculations and the results. We describe a detailed and concrete application of the PIQuant algorithm to a real biological network in the Results section.

Figure 1: **A simple influence network.** The network is composed of seven nodes and nine edges (a). The two paths (b,c) extracted from this network start from the source node **Ac** and end at the target node **Ph** (which denotes a phenotype of interest). The nodes **Ac** and **D** are annotated using experimental data, and have the values 2.0 and −1.0 respectively.




### BiNoM BioPAX utils & query

The BioPAX format was primarily conceived as a standard facilitating the exchange of data between various database systems [8]. As a consequence, this format was designed first to be machine-readable, but was not intended to be edited and modified by biologists. Furthermore, due to its adoption by large biological knowledge repositories, some BioPAX files can be really big, such as the *Homo sapiens* network from the reactome database [4] that has more than 6,000 reactions involving more than 8,000 chemical species (proteins, RNA molecules, metabolites, etc.).

BiNoM implements a set of functions precisely aiming at allowing end users to easily visualize and modify BioPAX files. The functions are using Java class introspection techniques to build a BioPAX class tree. Then, the content of the file can easily be accessed. For example, figure 6 shows all the information linked to the TRAIL protein, after a call to the BioPAX property editor function of the BiNoM menu has been made. Details are given in the BiNoM manual on how to display valid attributes, edit them, and how to visualize the complete BioPAX tree.

The BioPAX Query functions in BiNoM allow the user to work with huge BioPAX data files and extract the relevant information, by querying an index and retrieving data from it. The index corresponds to a mapping of the content of the BioPAX file on a labeled graph (an index file is created and saved, using the XGMML format). Various statistics can be displayed on the content of the index, such as the number of proteins, complexes, reactions, publications, etc. To start extracting relevant information, the user can query the index by gene name (and/or by any synonym of the gene) and start building a network centered around this molecule of interest. The extension of the network is done by adding different types of entities: complexes where the molecule of interest is involved, chemical species, reactions (with the possibility of including all

the sources and targets of the reactions) and related publications. Figure 7 illustrates an example of a small network extracted from the human apoptosis pathway downloaded from the Reactome database [4], and centered on the SMAC protein, with all the protein complexes in which this protein is involved and that were added using the BiNoM BioPAX Query functions.

### BiNoM Module manager

To facilitate the visualization of large molecular networks, we propose a set of functions that simplify them by creating modules from selected parts of the large network. This task, that we call *modularization*, is a semi-automatic procedure, where biological expert knowledge is used to assure the coherence of the newly created modules.

Most of the modules represent a detailed sequence of events that occur with a particular protein or protein complex, whose name can then be used to represent the whole module. This way, a simplified representation of a complex map can be produced, using the modules and their relationships as an abstracted version of the comprehensive network [18].

To facilitate the creation and management of modules, we have used in this version of BiNoM a new feature introduced in recent versions of Cytoscape (as of version 2.7) [22], known as *nested* networks. This feature allows to embed any cytoscape network in a (meta)node. It was introduced to allow the creation of network hierarchies and circular relationships. In BiNoM, we use this feature to facilitate the process of modularization of a large network. The BiNoM module manager integrates functions that allow to easily create a network from a list of subnetworks, packing individual nodes, merging different subnetworks, displaying information about metanodes and calculating the intersection between subnetworks.

### BiNoM Utilities

This set of functions corresponds to various small utilities that are not implemented in Cytoscape yet, but might be very useful for the analysis and manipulation of networks. For example, it is possible to automatically select all the edges between two nodes in the network in one operation, to generate the network corresponding to the double network differences between two networks A and B (creation of the two networks corresponding to $A - A \cap B$ and $B - A \cap B$), or update all the subnetworks of a session after some changes have been made to the initial one. The BiNoM Utilities also implement clipboard functions, giving the possibility to copy, add and paste selected nodes and edges and also to show the clipboard content.

## Results and Discussion

We propose to study a reaction network focusing on the transition from G1 phase (growth phase) to S phase (DNA replication phase) of the cell cycle [18] as an example of the use of BiNoM functions.

The network published in [18] describes in biochemical details the regulation of the well-known and charaterized tumor suppressor gene retinoblastoma (RB or RB1). The product of this gene operates at the heart of the cell cycle, acting as a signal transducer, connecting the cell cycle with the transcriptional machinery. The pathway in which RB is acting is disrupted in many human tumor types [28]. The comprehensive map of the RB/E2F network was built using CellDesigner [20]. It lists 80 proteins, 208 chemical species, 165 interactions, 176 genes, and recapitulates more than 350 publications, including information from different cellular types, thus making the map a generic map of the cell cycle regulation. It is composed of two main compartments: the cell, containing the cytoplasm, the nucleus and the nucleolus, in which the biochemical interactions such as association, dissociation, (de)phosphorylation, (de)acetylation, degradation, etc. take place; and the genes, which lists the target genes of the main transcription factors of the map, the E2F family members. A thorough description of the model, the methods used to build it and create simplified versions of it along with an interactive (clickable) map are available on our website (http://bioinfo-out.curie.fr/projects/rbpathway/interactive/rb_network.html/).

For the study presented here, we chose to concentrate on the G1 to S transition. We used the intersection of the 208 chemical species of RB/E2F network and the 280 chemical species listed in Reactome [4] for the G1-S transition (referred to as *Mitotic G1 G1/S phases* in Reactome). The resulting subnetwork contains 38 proteins, 98 chemical species, and 100 biochemical reactions (figure 8).

This map contains a lot of valuable information but rather difficult to extract. We present two ways to get some biological insight from this map, by using BiNoM functions. The first one consists of transforming the reaction network into an influence network in order to analyze experimental data on it. The second one implies a simplification of the comprehensive map by applying a method of reduction of the numerous interactions into modules without losing any content from the original map.

### Application of PIQuant on an influence network.

PIQuant algorithm can be used to perform a quantitative analysis on an influence network. In order to translate the G1/S reaction network into an influence network, we used a tool developed by the team of BIOCHAM [29] that is available online (http://contraintes.inria.fr/~soliman/cd2dot.html). To translate a reaction network into an influence network, the first one is pre-processed according to simple rules: (1)

BIOCHAM deletes all non-regulated degradation and syntheses reactions, (2) all intermediary chemical species with only one input and one output are suppressed, (3) if the reactions of synthesis and degradation of the chemical species deleted in (2) have distinct inputs and outputs, then these reactions can be merged, and (4) if they have the same chemical species as input/output, then the reaction is a reversible reaction and is replaced by a degradation [30]. A thorough description of the procedure together with an example of such a conversion is available in [31].

We applied PIQuant to the resulting influence network of the G1/S transition of the cell cycle (figure 9). We selected three target nodes as markers of the G1, S and M phases of the cell cycle. For the experimental data, we used expression data from a study of 57 bladder cancer tissue samples compared to 4 normal samples [32]. For each gene, the differential expression between tumor and normal tissue is assessed by a t-test. The t-test statistic value is used as the annotation for each gene. We selected the 19 nodes for which we had experimental data values as source nodes. Then, we constructed a text file listing nodes of the influence network and their annotation and we imported this file using the Cytoscape function "Import > Nodes attributes" in the Cytoscape session of the influence network. Figure 9 represents this influence network after its import.

PIQuant is applied to this network and its annotation, by using the function "Plugins > BiNoM 2.0 > BiNoM Analysis > Path Influence Quantification analysis". We selected the option "optimal and suboptimal shortest path" as the algorithm to extract the paths. The PIQuant score is then automatically calculated for each association between a source node and a target node. The user can browse the results on an interactive window detailing the different paths and their scores, and can also get a complete report, detailing the global and individual PIQuant scores from each annotated source node to each cell-cycle phase marker (for more details on the interactive window and the report, see the BiNoM manual). The global PIQuant score from each annotated source node to each target is represented as a heatmap on figure 10 (the list of nodes and all the PIQuant score values corresponding to the heatmap figure are available as supplementary table 1). We can see on this figure that most genes, in cancer cells compared to normal cells, influence positively the M and S phases (red coloring on the heatmap), indicating an enhanced proliferation for tumor cells. A clear difference can be observed when comparing the heapmap to figure 9, where the color values represent only experimental data values. The heatmap represents this time the integration of both experimental data and the network architecture through the PIQuant score.

### Modularization of the G1/S molecular map

The raw G1/S network is very detailed and may be hard to grasp at a first glance. To facilitate the analysis of the content, we propose to organize the reaction network as a modular network. The chemical species are clustered in groups, referred to as modules, in an semi-automatic manner, using BiNoM functions and biological knowledge. Each module represents in fact a sequence of events occuring with a particular protein. The modules are then linked by activating or inhibiting influences according to the information contained in the original diagram or derived from previous biological knowledge. A detailed tutorial on the construction of this modular network using BiNoM is described in the supplementary methods.

Briefly, we first decomposed the global network into its different components, by using name semantics to isolate the subnetworks in which each protein is involved (decomposition in material components). The 36 networks that are created this way may share a lot of common chemical species, so we went further by clustering the subnetworks having at least 25% of common chemical species. We renamed the 7 clusters obtained with a name that illustrates the content and the main function of the clusters (such as E2F1_RB, Wee1, etc.). Then, we checked the content of each module, making modifications if necessary by adding or deleting nodes, according to specific biological knowledge. For example, the module E2F1_RB is further decomposed in three different modules containing the proteins RB, E2F1 and E2F6. Finally, we generated a modular view of all the individual modules, using the Module Manager functions. BiNoM links the modules if they share components or edges. These edges are then interpreted as activation or inhibition by the user. Our final modular view is composed of 9 modules, with 22 edges connecting them (figure 11).

The modular view offers a simplified visualization of the complex network, without losing any information of the global map. The obtained model is more abstract but highlights some aspects that may not be evident from the comprehensive reaction network. For instance, it brings into relief feedbacks (positive, negative or feedforward) involving the major players of the cell cycle, and prepares the network for mathematical modeling. The translation of this modular network into a Boolean model is indeed straightforward. Another application for the modular model would be to analyze experimental data such as transcriptome or copy number variations (CGH). The "activity" of each module is based on the expression levels of the genes within the module, which can be visualized using a color code on the modular map. It's then fairly easy to analyze the difference between a disease and a normal state, or even to try to discriminate between different disease stages. We have produced such maps for the RB/E2F modular network to analyze bladder tumor samples, and we could observe a striking difference between the non-invasive and invasive states of the disease [18](the colored maps can be visualized at http://bioinfo-out.curie.fr/projects/rbpathway/case_study.html).

## Conclusions

Building a suitable model for systems and mathematical biology is a multi-step process, beginning with the collection of biological knowledge and progressing towards the formalization of a network and its translation in mathematical terms. BiNoM was designed to help during the intermediate steps of this process, by providing a convenient access to some standard Systems Biology representations such as BioPAX, by giving the possibility to manipulate the network by applying various algorithms (mostly based on graph theory) and map biological data to it. BiNoM is clearly not a tool for numerical simulations, but it provides functions to export final networks to the SBML and GINsim file formats (through the GINsim Cytoscape plugin for Boolean modeling), facilitating the import into various numerical simulators.

The current development of BiNoM includes novel functions covering various areas. A utility is under development that will allow the user to merge several independent CellDesigner maps into one by taking into account shared components. We are developing a new algorithm to find automatically minimal intervention sets to disrupt or modify the signaling flow for a given influence network (OCSANA algorithm, [33]). More precisely, an ensemble of paths can be defined for an influence network, between a set of input nodes and a set of output nodes. The algorithm is then finding what are the sets of nodes that will cut all the paths if they are removed from the network, a combinatorial problem is known as the minimal cut set or minimal intervention set. Finally, we are also working on an extension of BiNoM that will convert a CellDesigner molecular map to code corresponding to a web-based representations of biological networks using the Google Map API, with a possibility to easily navigate and zoom within the molecular map and curate the network through a dedicated web-blog (NaviCell software, http://navicell.curie.fr).

## Availability and requirements

- Project name: BiNoM
- Project home page: http://binom.curie.fr/ and http://apps.cytoscape.org/apps/binom
- Operating system(s): Platform independent
- Programming language: Java
- Other requirements: java 1.5 or higher, Cytoscape 2.7, 2.8
- License: GNU LGPL
- Any restriction to use by non-academics: none

## Acknowledgements

This work is supported by the APO-SYS EU FP7 project. EB, LC, DR, GS, EmB and AZ are members of the team "Computational Systems Biology of Cancer," Equipe labellisée par la Ligue Nationale Contre le Cancer.

## References


1. Adriaens M, Jaillard M, Waagmeester A, Coort S, Pico A, Evelo C: **The public road to high-quality curated biological pathways**. *Drug discovery today* 2008, **13**(19-20):856–862.
2. Cary M, Bader G, Sander C: **Pathway information for systems biology**. *FEBS letters* 2005, **579**(8):1815–1820.
3. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto encyclopedia of genes and genomes**. *Nucleic acids research* 1999, **27**:29–34.
4. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath G, Wu G, Matthews L, et al.: **Reactome: a knowledgebase of biological pathways**. *Nucleic acids research* 2005, **33**(suppl 1):D428.
5. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database**. *FEBS letters* 2002, **513**:135–140.
6. Saraiya P, North C, Duca K: **Visualizing biological pathways: requirements analysis, systems evaluation and research agenda**. *Information Visualization* 2005, **4**(3):191–205.
7. Gehlenborg N, O'Donoghue S, Baliga N, Goesmann A, Hibbs M, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, et al.: **Visualization of omics data for systems biology**. *Nature methods* 2010, **7**:S56–S68.
8. Demir E, Cary M, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, et al.: **The BioPAX community standard for pathway data sharing**. *Nature biotechnology* 2010, **28**(9):935–942.
9. Hucka M, Finney A, Sauro H, Bolouri H, Doyle J, Kitano H, Arkin A, Bornstein B, Bray D, Cornish-Bowden A, et al.: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics* 2003, **19**(4):524.
10. Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem M, Wimalaratne S, et al.: **The systems biology graphical notation**. *Nature biotechnology* 2009, **27**(8):735–741.
11. Bader G, Cary M, Sander C: **Pathguide: a pathway resource list**. *Nucleic Acids Research* 2006, **34**(suppl 1):D504–D506.
12. Karlebach G, Shamir R: **Modelling and analysis of gene regulatory networks**. *Nature Reviews Molecular Cell Biology* 2008, **9**(10):770–780.
13. Kitano H: **Systems biology: a brief overview**. *Science* 2002, **295**(5560):1662–1664.
14. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology**. *Annual review of genomics and human genetics* 2001, **2**:343–372.
15. Bauer-Mehren A, Furlong L, Sanz F: **Pathway databases and tools for their exploitation: benefits, current limitations and challenges**. *Molecular systems biology* 2009, **5**.
16. Oda K, Matsuoka Y, Funahashi A, Kitano H: **A comprehensive pathway map of epidermal growth factor receptor signaling**. *Molecular systems biology* 2005, **1**.
17. Oda K, Kitano H: **A comprehensive map of the toll-like receptor signaling network**. *Molecular Systems Biology* 2006, **2**.
18. Calzone L, Gelay A, Zinovyev A, Radvanyl F, Barillot E: **A comprehensive modular map of molecular interactions in RB/E2F pathway**. *Molecular systems biology* 2008, **4**.
19. Caron E, Ghosh S, Matsuoka Y, Ashton-Beaucage D, Therrien M, Lemieux S, Perreault C, Roux P, Kitano H: **A comprehensive map of the mTOR signaling network**. *Molecular systems biology* 2010, **6**.
20. Funahashi A, Tanimura N, Morohashi M, Kitano H: **CellDesigner: a process diagram editor for gene-regulatory and biochemical networks** 2003.
21. Zinovyev A, Viara E, Calzone L, Barillot E: **BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks**. *Bioinformatics* 2008, **24**(6):876.
22. Cline M, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, et al.: **Integration of biological networks and gene expression data using Cytoscape**. *NATURE PROTOCOLS-ELECTRONIC EDITION*- 2007, **2**(10):2366.

23. Tarjan R: **Depth-first search and linear graph algorithms**. In *Switching and Automata Theory, 1971., 12th Annual Symposium on*, IEEE 1972:114–121.

24. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J: **Graph structure in the web**. *Computer networks* 2000, **33**(1-6):309–320.

25. Novak B, Csikasz-Nagy A, Gyorffy B, Nasmyth K, Tyson J: **Model scenarios for evolution of the eukaryotic cell cycle**. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 1998, **353**(1378):2063.

26. Gleiss P, Stadler P, Wagner A, Fell D: **Relevant cycles in chemical reaction networks**. *Advances in complex systems* 2001, **4**(2/3):207–226.

27. Vismara P: **Union of all the minimum cycle bases of a graph**. *Electr. J. Comb* 1997, **4**:73–87.

28. Weinberg R, et al.: **The retinoblastoma protein and cell cycle control**. *Cell* 1995, **81**(3):323–330.

29. Calzone L, Fages F, Soliman S: **BIOCHAM: an environment for modeling biological systems and formalizing experimental knowledge**. *Bioinformatics* 2006, **22**(14):1805–1807.

30. Fage F, Soliman S: **From reaction models to influence graphs and back: a theorem**. *Formal Methods in Systems Biology FMSB'08, Lecture Notes in Computer Science* 2008, **5054**.

31. Calzone L, Chaouiya C, Remy E, Soliman S: **Qualitative modelling of the RB/E2F network, deliverable D3.1**. *ANR CALAMAR (ANR-08-SYSC-003)* 2011. [Https://tagc.univ-mrs.fr/welcome/IMG/pdf/livrable-3-1.pdf].

32. Stransky N, Vallot C, Reyal F, Bernard-Pierrot I, de Medina S, Segraves R, de Rycke Y, Elvin P, Cassidy A, Spraggon C, et al.: **Regional copy number–independent deregulation of transcription in cancer**. *Nature genetics* 2006, **38**(12):1386–1396.

33. Vera-Licona P, Bonnet E, Barillot E, Zinovyev A: **OCSANA: Optimal Cut Sets Algorithm for Network Analysis**. *In preparation* 2012.

## Figures

### Figure 1 - Visualization of the six BioPAX example files, provided in BioPAX 3.0 documentation.

The BioPAX 3.0 documentation available at http://biopax.org contains six simple examples of BioPAX 3.0 files that describe different aspects of biological network interactions (genetic interaction, short metabolic pathway, gene regulatory network, biochemical reaction, phosphorylation, protein interaction). Here we show how BiNoM visualizes these examples after their import. The BiNoM type of representation is indicated below the reaction type, in brackets (Reaction Network, Pathway Structure and Interaction Map). The graphical node and edge semantic is described in more details in the BiNoM manual.

### Figure 2 - Apoptosis pathway structure.

Representation of BioPAX data extracted from the Reactome database [4], corresponding to the Apoptosis pathway and imported through BiNoM, using Pathway Structure BioPAX representation. The green nodes represent pathways, the pink triangular nodes denote steps, while grey nodes indicate reactions.

**Figure 3 - Decomposition in material components.**

The two overlapping subnetworks found after the decomposition in material components of the cell cycle model of Novak et al. [25], corresponding to the components Cdc13 and Cdc2.

**Figure 4 - Decomposition in cycles.**

The figure shows two cycles for the CDC25 protein found after the decomposition of the cell cycle network model of Novak et al. [25].

**Figure 5 - A simple influence network.**

The network is composed of seven nodes and nine edges (a). The two paths (b,c) extracted from this network start from the source node **Ac** and end at the target node **Ph** (which denotes a phenotype of interest). The nodes **Ac** and **D** are annotated using experimental data, and have the values 2.0 and $-1.0$ respectively.

**Figure 6 - Extra information linked to the TRAIL protein.**

The information is automatically extracted from the BioPAX file upon import with the BiNoM I/O functions.

**Figure 7 - SMAC pathway subnetwork.**

Subnetwork extracted from the human Apoptosis pathway, starting with the SMAC protein (white square) and expanding to all protein complexes where this molecule is involved (grey squares) using the BiNoM Query functions.

**Figure 8 - G1/S network.**

Overview of the G1 to S transition network, corresponding to the intersection of RB/E2F network and G1/S network extracted from Reactome.

**Figure 9 - Annotated influence network.**

Influence network of the cell cycle G1/S transition generated by BIOCHAM. Colors represent differential expression obtained from transcriptomic data compared to normal tissue. Color intensities are proportional to the t-statistic values (red values indicate positive values corresponding to an activation, green values indicate negative values corresponding to an inhibition). The three grey nodes are markers for the different cell cycle phases: G1 (pRB_star), S (CDK2/Cyclin E1 complex phosphorylated), M (CDC2/Cyclin E1 complex phosphorylated).

**Figure 10 - Heat map representation of the PIQuant scores.**

The map shows the results of the PIQuant algorithm applied to the G1/S influence network. Color intensities are proportional to PIQuant score (red color indicates a positive value; i.e an activation, the green color indicates a negative value, i.e. an inhibition). Each line represents a source node while each column represents a cell-cycle phase phenotype: G1 (pRB_star), S (CDK2/Cyclin E1 complex phosphorylated), M (CDC2/Cyclin E1 complex phosphorylated).

**Figure 11 - Modular view of the G1/S network.**

Modular representation of the G1/S network, created by using a set of different BiNoM functions. Each node (pictured as a green octagon), represents a different module, or subnetwork. The edges connecting the modules represent the known influences between modules.

## Tables

**Table 1 - BiNoM path analysis algorithms.**

Short description of the three different algorithms implemented in BiNoM for path analysis. The "Directed paths search" toggles the search for directed or undirected paths. The "Finite radius" option let the user restrict the search to a given path length, in order to limit the size of the results and computation time.

| Algorithms | Directed paths search | Finite radius |
|---|---|---|
| Shortest paths (Dijkstra's algorithm | o | o |
| Optimal and suboptimal shortest paths | o | o |
| All non-intersecting paths | o | o |

**Table 2 - PIQuant score input data.**

Input data types description for the calculation of the PIQuant score.

| Data type | Description |
|---|---|
| Influence network (set of paths) | An influence network of interest composed of different species (proteins, complexes, Rna, small molecules), connected by edges representing activation or inhibition. A collection of paths will be extracted from the network, defined between input and output nodes of biological interest. |
| Experimental data | Experimental data related to processes described in the network. Species in the network can be annotated with experimental data values (consisting of a real number), such as an expression ratio or a t-test statistic value. |

## Additional Files

### Supplementary methods

Detailed tutorial for the creation of a modular view of the G1/S network using BiNoM functions.

### Supplementary table 1

PIQuant score values for all the source nodes (rows) and all the target nodes (columns) of the G1/S influence network.