

Saikat Chowdhury^{1,2} / Noopur Sinha^{1,2} / Piyali Ganguli^{1,2} / Rupa Bhowmick^{1,2} / Vidhi Singh¹ /
Sutanu Nandi^{1,2} / Ram Rup Sarkar^{3,4}

BIOPYDB: A Dynamic Human Cell Specific Biochemical Pathway Database with Advanced Computational Analyses Platform

¹ CSIR- National Chemical Laboratory, Chemical Engineering and Process Development Division, Pune, Maharashtra 411008, India. <http://orcid.org/0000-0002-0783-3959>.

² Academy of Scientific and Innovative Research (AcSIR), CSIR-NCL Campus, Pune, Maharashtra 411008, India. <http://orcid.org/0000-0002-0783-3959>.

³ CSIR- National Chemical Laboratory, Chemical Engineering and Process Development Division, Pune, Maharashtra 411008, India, E-mail: rr.sarkar@ncl.res.in. <http://orcid.org/0000-0001-7115-163X>.

⁴ Academy of Scientific and Innovative Research (AcSIR), CSIR-NCL Campus, Pune, Maharashtra 411008, India, E-mail: rr.sarkar@ncl.res.in. <http://orcid.org/0000-0001-7115-163X>.

Abstract:

BIOPYDB: BIOchemical PathwaY DataBase is developed as a manually curated, readily updatable, dynamic resource of human cell specific pathway information along with integrated computational platform to perform various pathway analyses. Presently, it comprises of 46 pathways, 3189 molecules, 5742 reactions and 6897 different types of diseases linked with pathway proteins, which are referred by 520 literatures and 17 other pathway databases. With its repertoire of biochemical pathway data, and computational tools for performing Topological, Logical and Dynamic analyses, BIOPYDB offers both the experimental and computational biologists to acquire a comprehensive understanding of signaling cascades in the cells. Automated pathway image reconstruction, cross referencing of pathway molecules and interactions with other databases and literature sources, complex search operations to extract information from other similar resources, integrated platform for pathway data sharing and computation, etc. are the novel and useful features included in this database to make it more acceptable and attractive to the users of pathway research communities. The RESTful API service is also made available to the advanced users and developers for accessing this database more conveniently through their own computer programmes.

Keywords: Computational Platform, Pathway Nomenclature and Ontology, Protein-Protein Interaction Data, Protein-Disease Mapping, RESTful API


DOI: 10.1515/jib-2017-0072

Received: August 14, 2017; **Revised:** January 15, 2018; **Accepted:** January 29, 2018

1 Introduction

Due to the advancements of several big data experiments and analyses, systems-level understanding of the functional mechanisms of biochemical pathways and their roles in governing multiple biological systems using state of the art molecular biology experiments is expected to expand at various directions in the near future [1], [2], [3], [4]. A vast amount of pathway-specific data from different experimental findings are now available in the published literatures and patents, and thought to be generated more in different aspects (i.e. in terms of volume and varieties) [4]. Applications of such data to address several puzzling questions of biological systems are presently considered to be of higher priority for the experimental and theoretical biologists. However, acquiring such highly dispersed data from public domains and collating them into a common hub of the biological knowledgebase for further analyses are one of the major challenges to the researchers [5]. Moreover, to keep a pace with the sustainable growth of the ongoing research and with the current findings from omics-based experiments conducted at various levels of the biological system it is indeed a challenge to the individual research group to keep up to date their research works with the vast amount of new information. On the other hand, annotation of newly identified biomolecular species (e.g. genes, proteins, RNA, miRNA etc.) and their

Ram Rup Sarkar is the corresponding author.

 ©2018, Saikat Chowdhury et al., published by De Gruyter.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.

functional roles for governing multiple processes require a common platform to be properly shared among the other scientific and academic communities for future experiments and references [6]. Most importantly, the recent developments of integrative approaches using bioinformatics tools to analyze big data generated from various Omics based studies have made these entire processes more easier than previous initiatives and thus a huge amount of data related to biochemical pathways are now available in literatures as well as in various other resources in raw or processed formats [5], [7], [8], [9].

2 Related Works

During the last few decades, data curators have made successful efforts to curate, annotate and visualize the detailed information of biochemical pathways in terms of different biological contexts [10], [11], [12], [13]. Most of the databases have made the curated data (i.e. networks, reaction mechanism, molecular information etc.) freely available in the public domain [5], [14]. However, the pathway related data, which requires a diverse set of information to represent its comprehensiveness, are still scattered and heterogeneously distributed in different databases in multiple formats [5]. Researchers working in this field have recently reviewed the current scenarios of these databases, in which a thorough comparison is performed to analyze the available data these databases serve to the users, their limitations in various aspects (i.e. data related and technical issues), and successively come up with the appropriate solutions which could help to improvise these databases [5], [14]. For example, the use of appropriate ontology, pathway nomenclature, heterogeneity in the information of same pathway data across multiple databases, absence of cross references of molecular interactions (i.e. Reaction ID, PPI ID, Literature references etc.), inability to define a standard boundary of pathway reconstruction, unavailability of the information of protein complex formation data (i.e. dimerization, trimerization, dissociation etc.), sub-cellular locations of the molecules and their translocation to various organelles within the cell during signaling events, unavailability of biological context specific (disease, tissue specific, mutation etc.) pathway information, absence of advanced computational tools for performing various pathway analyses studies, etc. are shown in these reviews as necessary and useful requirements to improvise the currently available databases [5], [14].

Moreover, hypothesis generation, followed by the mathematical formulations of biochemical reactions (i.e. model development) and its *in-silico* simulations using various mathematical approaches (e.g. topological, logical and dynamic etc.) to assess the roles of a biochemical pathway in governing different biological scenarios is also one of the major research interest to the computational biologists [5]. These types of mathematical simulations heavily depend on the data provided by the pathway databases and hence, the availability of computer readable pathway information (i.e. file systems, syntax, schema, reaction parameters, etc.) should also be considered carefully by the data curators [5], [14]. In this scenario, the importance of integration of various computational tools altogether with the database interface could be one of the major up-gradation of these databases from a simple pathway data sharing portal to a pathway data analyses platform [14]. Considering the current scenarios of multi-disciplinary research works in the fields of molecular and computational biology, such type of up-gradation would be always beneficial for a wide spectrum of database users. On the other hand, from the user's perspective, it is worth to mention that the database interface should be more users friendly and interactive through the manual as well as automated computer-guided operations [5], [14]. To make it possible, the use of advanced and useful database query language (e.g. SQL), appropriate file format (e.g. SBML), and API based web services would be much more effective and useful [5]. However, updating the database through proper annotations of the pathway molecules and reactions with the appropriate sub-cellular locations, and cross-referencing with external database sources and literatures are the major problems, which are currently faced by the developers of such databases [14]. As mentioned previously, the deluge of biochemical pathway specific data in various public domains of scientific literatures makes it almost difficult for the data curators to continuously update the data in their databases regularly. As a result, the pathway reconstructions and simultaneously illustrating the pathway images with the newer information during every update in the database are one of the major challenges faced by the data curators [14]. In summary, the requirements of the advanced platform for pathway data sharing process and the computational analyses tools with more user-friendly features for performing various computational tasks are the major demands, which could be included or modified in the existing or new pathway databases.

An initiative of developing a human cell-specific BIOchemical PathwaY DataBase, "BIOPYDB" is introduced here by taking into consideration of the current challenges discussed in the earlier section. After inserting minimal amount of curated raw data into the database, significant portions of the post-processing tasks are performed automatically in BIOPYDB. The raw data of this database are mainly extracted from the manual curation of data published in experimental articles and from the other popular databases (see Supplementary File 1). The post-processing operations (e.g. reconstruction and automated productions of the pathway images,

annotations and hyper-linking of pathway components and reactions with other databases, post-processing of computer readable files for pathway data sharing, mapping of protein molecules with different diseases, etc.) are then performed automatically with the help of BIOPYDB's in-built dynamic computational algorithms without any manual interventions. Data entry operations such as insertion, deletion, or modification of the pathway data are performed through SQL and the database is based on Relational Database Management System (RDBMS), which allow the entire operations in a more dynamic way.

Presently, it is providing the information of 46 different human cells specific, intra-cellular cell signaling pathways that are involved in various developmental events of cells and tissues, such as cellular growth, tumorigenesis, and immune cell activation, etc. A new pathway ontology and standard nomenclature system of cell signaling pathways are introduced in this database to allocate as well as index the curated pathways according to their biological functions and relevance. Each molecule and reaction of the pathways is automatically hyper-linked with various other resources for further references of the database users. Additionally, disease pathway (currently only Glioma specific pathway is available) is included in this database as a repertoire of biological context based (e.g. disease specific) human cell signaling pathway database. Also, the relationships between different proteins with various diseases (specific to human) are mapped and shown as a network of proteins and diseases. Furthermore, the architecture of the backend of this database is designed dynamically, which could be easily updated and modified after performing insertion, modification or deletions. Biochemical pathway related information such as images, networks, molecules and interactions list, protein-disease mapping, etc. shown in the database webpage are instantly generated from the data stored in the backend of this database and do not require continuous modification as well as manual changes in its frontends after any updates in the database. Such dynamic and automated process is specifically helpful to the pathway curators to populate the pathway information without giving any effort for further post-processing operations.

On the other hand, to develop this database as a platform for performing *in-silico* pathway analyses, useful mathematical tools, such as network or topological analyses of pathway networks (using graph theoretic analysis), logical analyses (using discrete time, semi-dynamic Boolean equations) and dynamic analyses (using ordinary differential equations) are made available through user friendly interface. The pathway data sharing and analyses platforms are brought together in this database into a single computational framework, which makes it easier for the user to perform multiple analyses on pathway data with less effort. To access the database from the external computational platform, RESTful API service is made available for the advanced users and software developers. Using this service, it will be possible to obtain the various types of pathway related data (e.g. list of all pathways, pathway description, pathway image, species, and reactions etc.) stored in this database.

In order to welcome the involvement of more number of data curators to collate pathway specific data for this database and to keep the database up to date with current experimental findings, BIOPYDB is also providing the facilities to upload new pathway data and simultaneously analyse it within a single platform. Based on the user's request and approval, the newly uploaded data will be verified by the curators of BIOPYDB and will be stored in the main database as a freely available data to the general users. Such crowd sourcing facility to populate the database will be very much useful to expand and update the database continuously in future. The database is now available in the public domain and common users can also suggest any changes in any of the existing pathway by providing their feedbacks/comments.

In the following sections, the detailed description of the pathway data; its pathway data collation techniques from different resources; implementations and operations of various pathway analyses tools; procedures of using pathway data upload system from the user end; and a brief discussion about its frontend (i.e. web interface) and backend (i.e. database schema and objects) are thoroughly discussed.

3 Database Architecture and Implementation

3.1 Pathway Nomenclature and Ontology

A standard nomenclature system with structured vocabulary is currently very much required for indexing different biochemical pathways in the respective databases [14]. It is observed that although the biochemical pathway databases have made a significant progress of archiving the pathway data in various modes and file formats, but they are not yet concordant with a specific pathway nomenclature system [14]. The naming convention of the signaling pathways used by the databases most often depend on the names of the ligands or receptor molecules (e.g. Hedgehog pathway or Notch pathway), and some cases it is based on the main target transcription factor of the pathway (e.g. NF- κ B signaling pathway). The ambiguity of naming the pathways is also a common factor of the databases, in which a particular pathway is named in multiple ways and then indexed into the database [14]. Hence, by following a consensus naming system with well defined vocabulary and a hierarchical tree of the pathway based ontology will be of great importance to the pathway data cura-

tors for indexing the pathways systematically and respectively searching, analyzing the pathway data more smoothly. The ontology tree defined by BioPortal for annotating the rat, mouse and human genes into pathway terms would be most appropriate [15].

Currently, the curators of BIOPYDB are using the similar pathway ontology tree defined by this well accepted and popular portal of biological ontology to index the signaling and disease related pathways into its database. However, to resolve the issues related to the structured vocabularies of naming the pathways, BIOPYDB introduces its own syntax and vocabulary. Currently, BIOPYDB provides 3 different functional (developmental, immunological and cell proliferation pathways) and 1 biological context dependent (disease pathways) nodes of signaling pathways, and has placed them at the top of the hierarchy in the BIOPYDB pathway ontology tree. These top most level nodes are further linked with the child nodes in its down-stream branches. For example, under the node of immune signaling pathways, the cytokine signaling pathways constitute a separate sub-category (child node), which is further classified into distinct interleukin families like IL-2 family, IL-10 family, etc. Each of these interleukin families is further assigned to different interleukin signaling cascades, which are having similar functionalities. More number of different categories will be included in the subsequent updates of this database. To assign the names of the new pathways and make them distinct from each other within a same family and sub-family, the pathways are named according to either ligand (e.g. IL2, IL4) or the main receptor protein (e.g. TLR, T-cell receptor or TCR) molecule, which triggers that particular signaling cascade. The following vocabularies (*a*, *b*, *c*) are used for the nomenclature of the signaling pathways involved in various cellular functions (*a*, *b*) and disease pathogenesis (*c*):

- a. X Ligand(s) Stimulated Signaling Pathway(s)
- b. Y Receptor(s) Mediated Signaling Pathway(s)
- c. Deregulated Signaling Pathways in Z

where, *X*, *Y*, and *Z* represent the name of the Ligand or the family of the Ligands, Single or a family of Receptors molecule, and name of the Disease respectively.

If there are multiple ligands/receptors associated with the signaling pathway, then the logical operator "AND" is used in the vocabularies ("*a*" and "*b*") to separately mention the names of each of the ligands/receptor molecules. Introduction of such detail classification and nomenclature system is useful to the data curators as well as the users for searching/browsing a specific class/family of pathway(s) in the database. It is also helpful to understand and compare the different pathways activated/stimulated under different extra-cellular stimuli. The information about the context specific pathways (e.g. the disease pathways) is also indexed here by using a controlled vocabulary (*c*), which could be further used to differentiate the deregulated or malfunctioned pathways from its normal counterparts and their involvements into a particular disease (see Supplementary File 1). A flowchart describing the entire process of implementing and processing the pathway nomenclature system in the BIOPYDB database is depicted in Figure 1.

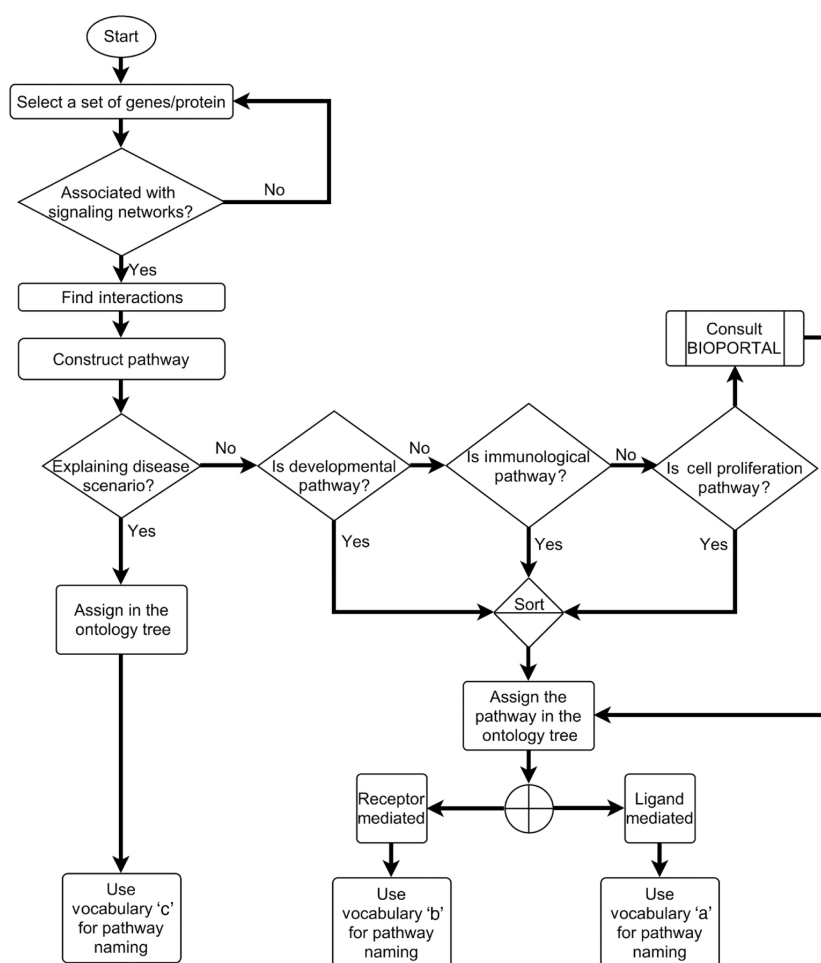


Figure 1: Flowchart describing the pathway nomenclature system used by the curators of BIOPYDB database.

3.2 Pathway Data Curation and Processing

Each of the interactions/connections/molecular reactions represented in the pathway data/image is acquired from published articles (around 520 articles are manually searched/referred for cross validation of interactions) to provide the appropriate experimental references. One of the main objectives of developing this database is to make a resource of biochemical pathways by feeding minimal information. Hence, to populate the database by feeding optimal amount of data, at first the pathway curators of BIOPYDB have consulted various similar resources (see Table S1 of Supplementary File 1) to gain the basic topology of each of the signaling networks. Most often, the topology and the pathway related data (i.e. number of molecules, interactions etc.) found in the external resources are highly heterogeneous (in terms of pathway names, sub-cellular locations of the pathway species, total number of interacting pathway components etc.) and hence it is required to normalize these piecewise information/data followed by its validation through literature reference, and/or cross-referencing with Protein-Protein Interactions (PPI) or Disease specific (e.g. cancer) databases [5], [14]. Figure 2 represents a schematic diagram of the different resources of pathway data that are used by BIOPYDB pathway data curators to get the information about the basic topology or the core functioning module of the pathway of interest. Later, the data collated from these resources are expanded and reconstructed by including more number of pathway species (inorganic or organic molecules, proteins, lipid, carbohydrate, ions etc.), cross talks, chemical or physical properties of the interactions, etc., which are found to be involved in that particular pathway, as recorded in the literature references, but are not included in any existing pathway databases. The pathways become more up to date, authentic and accurate for the users after inclusion of such comprehensive information from different literature sources.

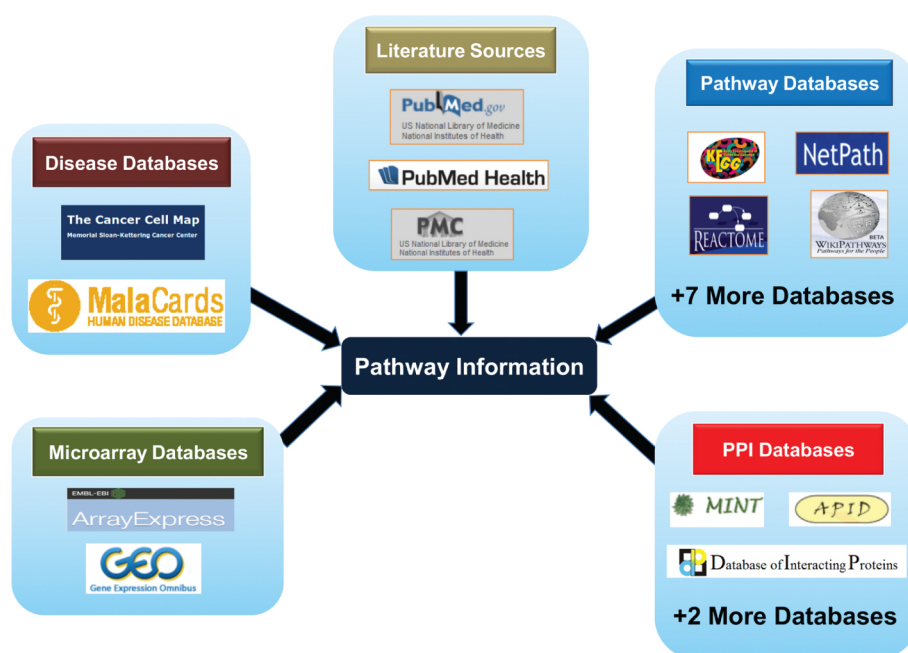


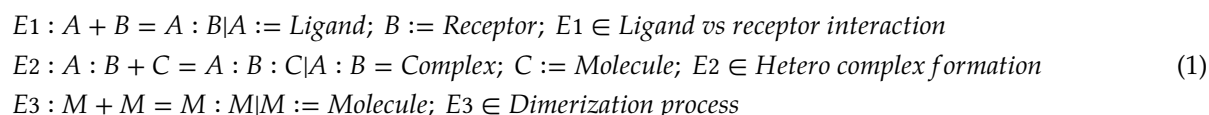
Figure 2: External data sources used to collect pathway information for BIOPYDB

In order to build or update the pathways, BIOPYDB uses a specific protocol (see Supplementary File 1) for further pathway data normalization and updating with the latest information and new experimental findings into its database. The following subsections provide a brief description of the protocol used for preparing the BIOPYDB database.

3.3 Processing of Molecules Table

Followed by the nomenclature and ontological assignment, the core molecules (i.e. proteins, genes, metabolites, RNA etc.) involve in a particular pathway of interest are searched in literatures and other databases. Each molecule is further categorized according to its molecular and chemical properties (e.g. Protein, carbohydrate, Gene, Ion, RNA etc.), and is then indexed in the database by its corresponding BIOPYDB Pathway ID. An abbreviated name, which is used in the published literatures or UNIPROT database, is assigned to each molecule. Also, the information of the sub-cellular location of the molecule is extracted from the published literatures and other databases. Moreover, it should be noted that, if multiple sub-cellular locations (e.g. cytoplasm and nucleus) are found to be associated with the same pathway molecule, prefix “Nuc_” is used to annotate the nuclear location to distinguish the pathway species with its cytoplasmic counterpart. No separate prefix is used for cytoplasmic species. For example, GLI1 (an important protein in Hedgehog pathway) is found in the cytoplasm as inactive at the time of no stimuli from hedgehog ligands but is found to be active and translocate into the nucleus followed by the activation of hedgehog signaling event. In this case, GLI1 protein in Hedgehog pathway is named as GLI1 and NUC_GLI1 for its cytoplasmic and nuclear counterparts respectively.

Furthermore, to show the molecular complexes formed by proteins and other molecules in the pathway, the following vocabulary is used. For example, let us consider the following reactions (i.e. E_1 ; E_2 ; E_3) from a signaling event in which the complexes are formed as products.



Different types of complexes formed in the signaling cascades by various chemical reactions are indexed in the molecule list by using the vocabularies $A:B$; $A:B:C$, and $M:M$ etc.

3.4 Hyper-linking and Annotation of Pathway Molecules

Hyper-linking and annotating of each protein molecule for all the pathways in a database is a time-consuming process, as these require more manual interventions and efforts. In BIOPYDB, minimal pathway information is

required to provide from the pathway curator's end, and its in-built pathway update engine running at the back-end automatically processes the hyperlinks of each of the protein molecule with other popular databases. For example, to get the information of the corresponding gene and amino acid sequences of a particular protein, BIOPYDB automatically hyper-links the protein molecule with NCBI-gene and UNIPROT databases respectively. Other popular databases *viz.* KEGG [16], WIKIPATHWAYS [17] and HUMANCYC [18] are also hyper-linked with the protein molecules to make this database more useful and authentic. To get the information of other interacting protein molecules against a particular protein of interest, each protein molecule of a pathway is provided with the interactive access of protein-protein interaction databases *viz.* PIPs [19] and STRING [20] databases. Also, to search the diseases associated with each of the protein molecules in a pathway, BIOPYDB provides the hyperlink with GeneCards [21] database, which is a very useful database for acquiring information of protein and disease relationships. The tissue-specific expression of a gene/protein is also an useful information, which is also served in BIOPYDB by providing the hyper-links of the gene/protein molecules with a popular database: TiGER [22] for this purpose. These automated processes of cross-linking the protein molecules with other relevant and popular databases are based on its in-build pathway update engine written in PHP, Perl, and Python languages.

3.5 Processing of Molecular Interactions Table

The molecular interactions or reactions present in each pathway in BIOPYDB are manually curated from the experimental data available in literatures and are successively stored in a master table called "Interactions". To include a new interaction in this table, the evidence of the interaction should be supported by at least one experimental reference from a published article. Any new interaction, which qualifies this criterion, is then included in the "Interactions" table with its corresponding BIOPYDB pathway ID. The corresponding literature reference of each interaction is then hyper-linked with the associated Pubmed ID of the article. The molecular and chemical properties (e.g. phosphorylation, ubiquitination, dimerization etc.) of different types of biochemical reactions are tabulated with each of the interaction in the "Interactions" table. The sub-cellular location, (e.g. extracellular space, membrane, cytoplasm etc.) in which the interactions are found to occur, is also tabulated with each interaction in this table. To maintain the direction of the reaction cascades, each of the reactions is considered as the directed edges connecting a source molecule with another target molecule (i.e. protein, complex, ions etc.). The source molecule (represented by "Protein A") is the upstream molecule of the target molecule (represented by Protein B) in the "Interactions" table. All such molecular interactions of a pathway are automatically hyper-linked with iRefIndex database [23], which provides a standardized indexing of the non-redundant molecular interactions by taking the popular protein-protein and molecular interactions databases such as BioGRID [24], IntAct [25], HPRD [26] etc., into consideration. Mapping the interactions with the iRefIndex database will be very much useful for the users to verify as well as collate the information of that particular interaction available in the other popular PPI databases.

3.6 Processing of Disease, Reference and Mathematical Model Tables

Deregulations in the different components of the biochemical pathways and their correlation with several disease pathogenesis are now one of the major research interests to the clinical and experimental biologists. To get such information flawlessly and with less effort, BIOPYDB is providing an interface to search and fetch the protein-disease mapping data from MalaCards: Human Disease Database [27]. The disease data is fetched from this database for each protein mapped with the particular BIOPYDB pathway ID. The diseases mapped with each protein are then tabulated in BIOPYDB web interface and a network picture of protein-disease map is provided for better visualization and further analyses. It should be noted that the protein-disease data provided in BIOPYDB is solely owned by the MalaCards database, and BIOPYDB is providing an advanced, automated interface to fetch and process the data in a more user-friendly process.

BIOPYDB also maintains a dedicated database of all the relevant references or published literatures related to each pathway in "Reference" table. The reference table is generated automatically by fetching the citation information from Pubmed database. The users can get this entire "Reference" list by accessing the pathway-browsing interface of BIOPYDB. Thus the literature mining for getting the information of a particular pathway will be much easier to the experimental biologists. Furthermore, to ease the efforts of theoretical biologists for obtaining information of previously published mathematical models developed for different biochemical pathways, BIOPYDB has also made a database of the literature references related to the mathematical models (e.g. Graph theoretical, Boolean, Ordinary or Partial Differential equations etc.) and tabulated in the "Model" table. The Pubmed id or DOI is also provided for each model and their hyper-links are automatically generated.

3.7 Processing of Pathway Image and Textual Data

There are various file formats used in BIOPYDB for sharing and distributing the pathway data to the users. Users can download the data without any restrictions and charges. For commercial uses, the users are required to take appropriate license from the BIOPYDB developers. The entire process of pathway data processing is described in Figure 3 for better clarification. The pathway data provided by this database is broadly categorized into two categories: (i) "Pathway Image" and (ii) "Textual Data". Both are automatically processed from the BIOPYDB main data sources (i.e. Molecules list, Interactions list etc.) for further sharing and the downloading purposes. The pathway images are processed in JPEG, SVG, and PNG formats. Two types of pathway images are produced *viz.* *Structured or Hierarchical* and *Unstructured or Network*. The Structured pathway image is produced by allocating the pathway species in a hierarchical fashion, in which the ligands and receptor molecules (present in the extracellular and membrane regions) of a pathway are placed at the left (in left to right hierarchy) or top (in top to bottom hierarchy) in the pathway image. The pathway components found in the cytoplasm and in nucleus are placed accordingly in the subsequent hierarchical levels (i.e. Extracellular region, Membrane, Cytoplasm, Nucleus) respectively. Another hierarchical location "Output" is introduced in the pathway image to signify the target genes/proteins produced at the end of the signaling cascades. The pathway molecules belong to the hierarchical location "Output" are placed at the rightmost or at the bottom of the *Structured or Hierarchical* image. Though this region does not indicate any physical sub-cellular location, but to distinguish the pathway molecules from the target proteins, inclusion of such location in the pathway image will be very much useful for better simplification. The sub-cellular locations, required to maintain this hierarchy, are fetched from the "Molecules" table generated at the beginning of the pathway curation. On the other hand, the connections/edges between two pathway molecules are fetched from the "Interactions" list. Different color codes and shapes are used to classify different types of molecular interactions enlisted in this table. Thus, an *in-silico* structured pathway image, resembling the internal cellular environment is reproduced automatically in the database interface. The unstructured pathway image (i.e. network) is also generated to analyze the topological properties of the biochemical signaling pathways. Such type of network is viewed as a directed graph in which the direction of the interaction between the two species of a pathway (i.e. Source and Target) is generated from the information provided in the "Interactions" table. Both of these pathway images (i.e. hierarchical and network) are generated automatically by using the popular graph drawing software: Graphviz [28] and the in-house codes written in PHP.

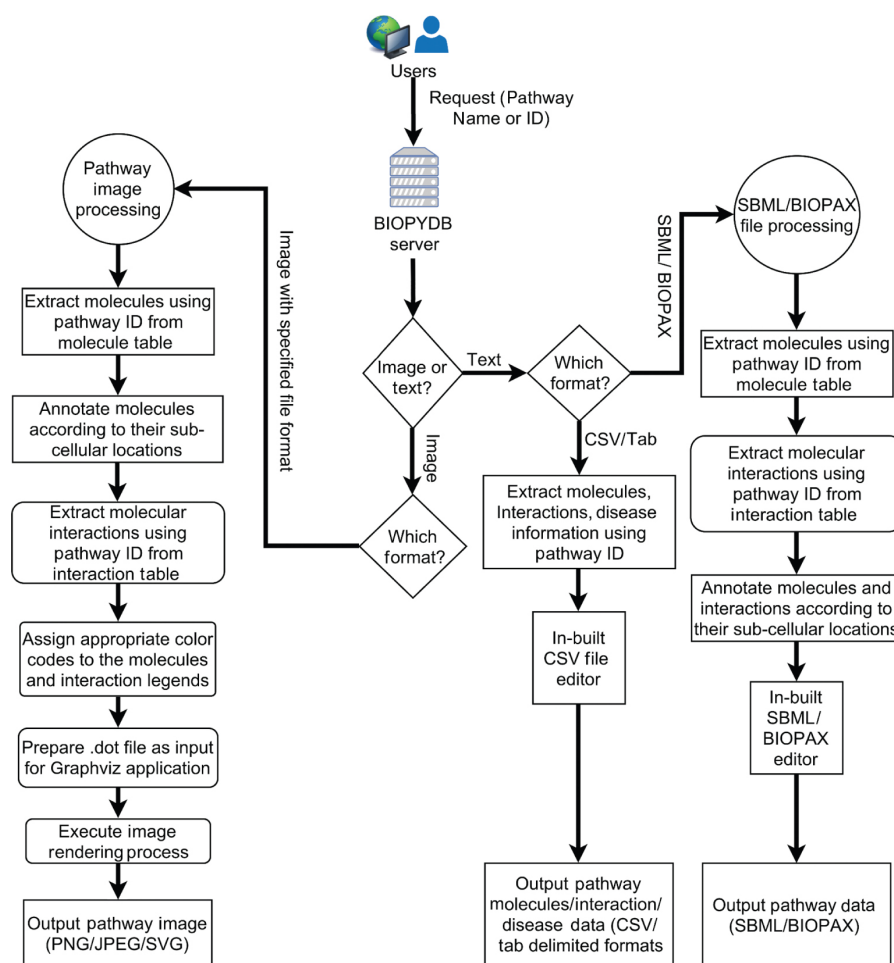


Figure 3: Flowchart depicting the process of pathway image construction and textual data preparation in the back-end of database server.

On the other hand, the textual data (i.e. information of molecule, interactions, diseases etc.) of biochemical pathways are generated in BIOPYDB through an automated fashion by fetching the raw data from the database tables' viz. "Molecules", "Interactions", and "Diseases". The textual data for each pathway are processed in simple flat files i.e. either in TXT or CSV formats or in XML based SBML (Systems Biology Mark-up Language) and BIOPAX (Biological Pathway Exchange) formats [29], [30]. SBML and BIOPAX files are computer-readable, vocabulary based, widely used standard pathway sharing file formats, which are useful for pathway visualization and performing mathematical modeling. The dynamic engine running on the backend of BIOPYDB web services automatically processes these varieties types of textual data. To generate runtime BIOPAX format of the biochemical pathways, BIOPYDB uses the open source, third-party tool Sig2BioPAX for converting the biochemical reactions into BioPAX level 3, OWL format [31]. Using such automated process of generating pathway images and textual data, the data in BIOPYDB can be readily updated and thus does not depend on any manual intervention by the database administrators and curators.

3.8 Data Storage System

Relational Database Management System (RDBMS) is used in BIOPYDB for managing its entire database. MySQL database server is used here for this purpose from which the stored data can be accessed through Structured Query Language (SQL). Different types of pathway related data (i.e. molecules, interactions, diseases etc.) are stored in different "Tables". These tables are the logical objects of the relational database within which the relations are established through unique pathway ids or primary key [32]. The structure of the database is made very simple and can be easily understood. BIOPYDB stores the pathway specific data (within tables) in such a way, so that after inserting or updating the pathway data, the successive outputs of pathway images, annotation and hyper-linking of pathway molecules with other databases, mapping the molecular interactions with iRefIndex database, or generating the textual data for pathway data sharing etc. will be automatically mapped by the corresponding unique pathway IDs and successively stored in the respective tables of the database. This automated process and the structure-function relationship of BIOPYDB are presented more precisely in Figure

4. In this figure, the entire data structure and its relationship with the available features are presented in a simple relational schematic diagram. To understand the database architecture and its interaction procedures with the end-users, a detail description of this diagram is provided in the Supplementary File 1.

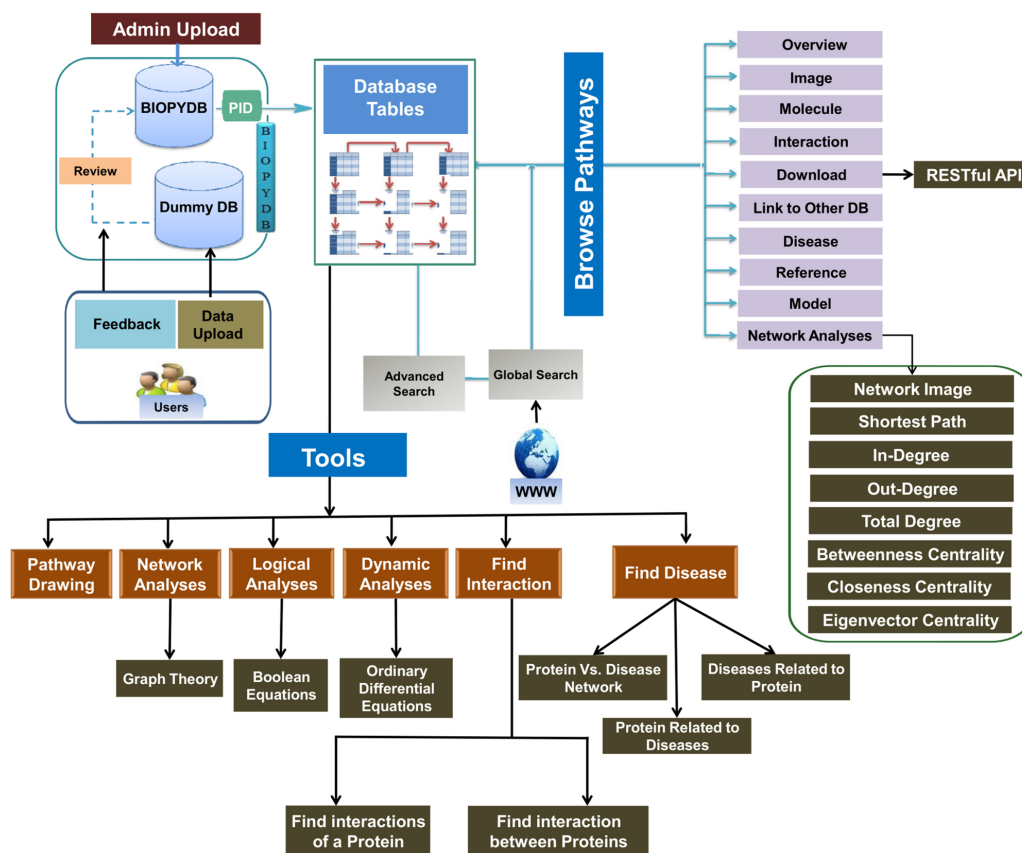


Figure 4: Schematic diagram of BIOPYDB database architecture.

In the backend of this database, it contains the logical objects or “Tables”, which individually store the pathway related data such as Pathway Name, Pathway Info, Overview of each of the pathway, the cross-talks of each pathway, Molecules, Interactions, Diseases, information about the developed mathematical models, References etc. All the tables contain a primary key (Pathway ID), which is used to connect each of the tables while performing the data searching and entry operations in the database. Hence, each and every data in the database stored in the database tables (pathway molecules, interactions, references etc.) are associated with such unique pathway id, which makes the different operations (e.g. insertion, deletion, modification etc.) in the database more accurately using structured query language (SQL) with the help of any dynamic programming language such as PHP. This schema provides this database more fluidity to perform post-processing tasks in an automated fashion by consuming less time and manual interventions.

4 Application

4.1 Resource of Biochemical Pathways

BIOPYDB is currently providing the biochemical, intra-cellular signaling pathway data under four different categories *viz.* (i) *Developmental Pathways*, (ii) *Immunological Pathways*, (iii) *Cell Proliferation Pathways*, and (iv) *Disease Pathway*. In the *Developmental Pathways* category, two important developmental pathways *viz.* Hedgehog and Notch pathways are included. Hedgehog pathway is further sub-categorized into three different pathways based on the three homologues of Hedgehog ligands (i.e. Sonic, Desert and Indian Hedgehog). Notch Pathway is also divided into two pathways which are activated by either JAGGED or DELTA ligands. Moreover, BIOPYDB has a rich collection of cell growth regulator pathways eminently represented by the Epidermal Growth Factor Receptor (EGFR/ErbB) mediated Mitogen Activated Protein Kinase (MAPK) pathways under the “*Cell Proliferation Pathways*” category. The EGFR/ErbB mediated signaling system forms a family of pathways, which is classified here on the basis of ligand-receptor interaction. Apart from the known Growth fac-

tors, like the Epidermal Growth Factor (EGF), heparin-binding EGF (HB-EGF) and Neuregulins (NRG1, NRG2, NRG3, and NRG4), a few of the lesser known ligands of the family like Amphiregulin, Betacellulin, Epiregulin are also included. Strikingly, Transforming Growth Factor (TGF- α), a TGF family ligand, due to its structural similarity with the growth factors of EGFR/ErbB family, is classified into the EGFR/ErbB family of signaling. A unique ligand of neuregulin subfamily, Neu Differentiation Factor (NDF- β), is also included in this category. In contrast to the other NDF isoforms, which generally trans-activate ErbB-3 co-expressed with ErbB-2, NDF- β can also trans-activate ErbB-3 co-expressed with ErbB-1. BIOPYDB also contains an exhaustive list of “*Immunological Pathways*” that includes pathways related to both innate (e.g. TLR pathways) and adaptive immunity (e.g. T cell signaling pathways). It contains a repertoire of cytokine signaling pathways as well, that includes different interleukins, interferons, tumor necrosis factors, tumor growth factors and colony stimulating factors, all of which plays a pivotal role in regulating the immune-regulatory network. To provide the context-specific pathways, BIOPYDB also provides the signaling pathways, which are found to be deregulated in a particular disease scenario under “*Disease Pathways*” category. The malfunctioned signaling pathways, which are only stimulated by PDGF, IGF1 and EGF ligands in Glioma, are currently provided in this section. Later, it also intends to provide all other possible pathways, which are responsible for the development of Glioma in human. All these categories of pathways are easily accessible from the “Browse Pathways” tab provided in the home page of BIOPYDB.

To access the information of the pathway of interest, the users are required to click on the corresponding link of the pathways in the pathway browse section. A small description of each pathway, including the pathway statistics (number of molecules, interactions, articles, diseases etc.), will be available in the “Overview” section of the pathway data. Besides, the phenotypic expressions (e.g. cell division, cell growth, apoptosis etc.) and cross-talks with other pathways with proper literature references will be also available in this section for each pathway.

Presently, BIOPYDB contains 3189 molecules (Proteins, Protein complexes, Inorganic molecules, Mutated proteins, Secondary messengers, Phospholipids and Lipid molecules) and 5742 molecular interactions or connections of 46 manually curated pathways. For accuracy and authenticity of the pathway data, each and every interaction included in the pathway is cited (by a published article with its corresponding Pubmed ID). Users can cross verify the pathway data immediately by accessing the hyper-links provided in the “Interactions” tab under the “Browse Pathway” option. It also facilitates to cross-check the pathway data in similar 17 different databases by automatically searching the links in these databases and thus helps to connect the users with these resources immediately.

4.2 Resource of Proteins/Genes Involved in Biochemical Pathways

BIOPYDB can also be used as a resource of signaling protein(s)/gene(s), which is evidenced in the literature regarding direct or indirect influence in the intracellular signal transduction cascades. Currently, it contains 2748 such proteins/genes, which are easily accessible or can be searched through the general “Search” as well as “Advanced Search” option provided in “Home” page. The query result gives a short description of the protein, as well as its genetic and amino acid sequence data from NCBI-GENE and UNIPROT respectively [33], [34]. Users can also get the links of the proteins from other pathway resources like KEGG, WIKIPATHWAYS and HUMANCYC databases [16], [17], [18]. Besides, it provides the hyper-links of the protein-protein interaction databases from the databases PIP; Disease related data from GENECARD and tissue-specific expression pattern data from TiGER database [13], [19], [22].

4.3 Resource of Protein-Protein Interaction Data

Protein-protein interaction data with network representation and proper references are also available through the “Find Interaction” application, available in “Tools” section. The type of interactions (i.e. chemical nature) of each interaction is also available with each of the interaction. Currently, BIOPYDB has included 25 different types of reactions/ interactions/ connections in the signaling networks. The types of interactions are *Cholesterol Modification, Physical interaction, Inhibition, Phosphorylation, Activation, Nuclear Translocation, Transcriptional Co-repression, Protein Production, Transcriptional Co-activation, Protein Recruitment, Auto-phosphorylation, Ubiquitination, Transcriptional activation, Complex formation, Stimulation, Dissociation, Dephosphorylation, Homodimerization, Heterodimerization, Deubiquitylation, Calcium exchange, Acetylation, Phospholypase reaction, Proteolytic cleavage, and Enzymatic Reaction* [35], [36], [37], [38], [39], [40], [41], [42]. Information related to the type of chemical reactions or interactions occurring between two pathway molecules present in a pathway will help the users of BIOPYDB to understand the biochemical processes, which are regulating the signaling cascades inside the cell. Moreover,

to further cross verify the interaction data, each interaction in the database are mapped and hyper-linked with the iRefIndex database [23], which consists non-redundant molecular interaction data collated from different protein-protein interaction databases.

4.4 Resource for Visualizing the Protein-Disease Mapping

It is well known that certain malfunctions caused by the mutation of proteins in biochemical pathways could be responsible for various types of diseases in human body [43], [44], [45]. Therefore, it is worth to mention that protein-disease mapping of the protein molecules in the signaling pathways is also important to understand the importance of that pathway in disease pathology. To include this feature into the database, BIOPYDB provides an interactive interface to dynamically fetch the information of various diseases from protein-disease mapping database: MalaCards [27] and displays the results in a more user-friendly way. All the information provided in this section is solely owned by MalaCards database and the developers of BIOPYDB do not hold any claim related to the protein-disease mapping data. It simply provides a resource for fetching and visualizing protein-disease network as image and tabular formats in the BIOPYDB interface. The diseases associated with each protein of the pathway of interest are automatically hyper-linked with the MalaCards database ID. Currently, BIOPYDB has dynamically mapped around 6897 diseases with the proteins present in the pathways in BIOPYDB database. The disease related information is available through the “Disease” tab under the “Browse Pathway” option. It is also available through the “Find Disease” application available in the “Tools” Section.

4.5 Computational Platform for Pathway Data Analyses

One of the unique features of BIOPYDB is its in-built pathway data analyses platform for performing network (using graph theoretic analysis), logical (using discrete time, semi-dynamic Boolean equations) and dynamic (using ordinary differential equations or ODEs) analyses of biochemical pathways [46], [47], [48]. No other similar resources/databases provide this wide ranges of computational tools all together in a single platform to the common users. Moreover, to make this database more interactive and user-friendly for the large community of experimental and theoretical biologists, developers of BIOPYDB have included various technical features and applications to search, retrieve, annotate or analyse the data stored in the database. The entire resource is easily searchable and does not require any separate hardware or software installation in the user’s local machine. To view the pathway image and network, users are required to use an SVG compatible web browser, which is now available in any modern web browsers, like Internet Explorer (version >9), Mozilla, and Google Chrome etc. A schematic diagram showcasing all the available tools and the stepwise guidelines of their operational protocol is presented in Figure 5. An additional data file containing a brief description of all the technical features available in BIOPYDB is provided in Table S2 of “Supplementary File 1”. Following is the brief descriptions of the technical or computational features available in BIOPYDB.

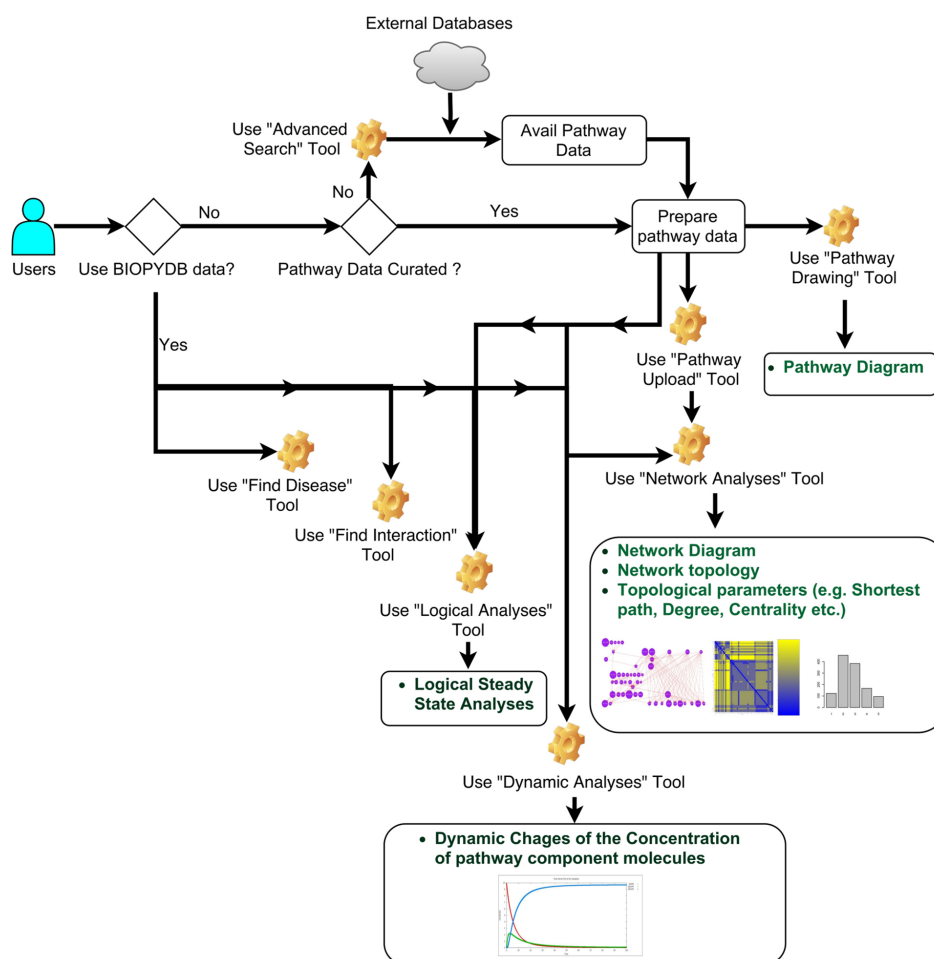


Figure 5: Schematic diagram showcasing all the available tools and their working protocols.

4.5.1 Searching/Browsing Pathway Data

The text search option is available in the home page of BIOPYDB. The search box includes “auto-suggestion”, which suggests the pathway names, proteins or diseases from the database while typing into the search box and thus expedites and eases the text searching procedures. Users can also browse and select a pathway of interest by using “Browse Pathway” option. The “Browse Pathway” option has two different tabs for browsing the pathways “By Name” or “By Category”. After selecting a pathway from any of these options, users will get the pathway details which are provided in 10 different tabs: *Overview, Image, Molecule, Interaction, Download, Links to Other DB, Disease, References, Model* and finally *Network Analysis*. Users can click on any tab and get the relevant information about the pathway.

4.5.2 Advanced Search Option

The advanced search option is provided to the users to execute complex and multiple queries through this database. Search queries, such as pathway name; drugs or inhibitors of a pathway; disease caused by the mutation of proteins; articles searching operations using Pubmed IDs, E.C. number, etc can be executed in this section. Using this unique application of BIOPYDB, users not only can search in BIOPYDB database, but also can perform search queries to the other 17 similar signaling pathway resources, as well as in PUBMED, ExPASy, BRENDA, METACYC, and GENECARDS databases. This unique advanced search query options will make this database more attractive to the users for collating diverse sets of biological data from a common web application platform.

4.5.3 Pathway Image Drawing

This application is useful for automated pathway drawing purposes, where users do not have to manually draw the pathway diagram; instead, it can automatically annotate and draw the molecular connections between the pathway components based on the information provided by the users. Simultaneously, it renders the pathway image in the SVG format and hosts it into the web browser. To use this application, users are required to simply insert the molecular entities and their binary interaction or connections in the specified fields mentioned in the web interface of this tool. Users can also specify the different types of inserted molecular entities (e.g. Protein, Protein complexes, Secondary messenger, Inorganic molecule, Mutated protein and Phospholipid), its sub-cellular locations in the cell (e.g. Extracellular region, membrane, cytoplasm, nucleus) or can allocate the molecule in “output” section by using the simple drop-down list available in the pathway data upload form. Similarly, in the interaction section, total 25 different types of molecular reactions are enlisted and can be used by using the drop-down menu. For better visualization and image rendering, users are required to provide at least four molecules from four different sub-cellular locations and their interactions in the specified fields. Users can add or delete any molecules or the reaction in this dynamic web page form. After filling up all the information, the user has to click on the “Draw” button, which will instantly generate a pathway image in an SVG format and the pathway image will appear in the web browser. Here, the different types of molecule and interactions will appear in different colors according to their molecular properties and thus it helps to differentiate the reaction cascades more appropriately in the newly generated image.

4.5.4 Network Analyses

This application allows the users to perform topological analyses of the biochemical pathways using graph theoretical method [46], [47], [49]. This mathematical analysis can be performed on the pathways, which are already available in the BIOPYDB database. Using this tool, the user can calculate various network parameters, such as *Connectivity parameters* (all pairs shortest paths, in-degree, out-degree, total degree and their corresponding distributions), and *Centrality parameters* (eigenvector, betweenness and closeness centrality of a pathway network). All pairs shortest paths (of each molecule of the signaling pathway) are presented by a matrix heat map, and the other parameters values are shown by bar diagrams. To execute this application, first the users are required to select the pathway to view the list of the binary interactions of the pathway molecules on which the topological analyses will be performed. After that, users can directly run the graph theoretical simulation on the provided binary interactions or can add or delete any pathway species and its corresponding interactions in the provided list and then can run the simulation. The inclusion of this feature into the user-interface makes it a useful platform to perform the protein knock out experiments in a signaling network, where one can simultaneously observe the variations of the different network parameters. Each simulation will provide the network picture, average values of the network parameters, and the bar plots of each parameter values for each pathway molecule of the network. Users can also download the network picture and bar plots in SVG, PNG, PDF or JPG formats.

4.5.5 Logical Analyses

Users can also perform logical (i.e. semi-dynamic, discrete time dynamics) analyses of the expression/activity pattern of molecular entities of a pathway using the tool provided in this section. It is based on the in-built “Boolean Analysis” simulation technique widely used for analyzing the activity (i.e. up-regulation and down-regulation) of pathway molecules in the biochemical and gene regulatory networks [46], [50], [51], [52], [53]. This application has three sub-sections, mainly divided for three different purposes. The first section allows the users to perform Boolean analysis of the pathways, which are present in BIOPYDB database. The second sub-section allows the users to perform Boolean analysis of a new (user-defined) pathway and the third sub-section is used for revisiting the previous Boolean analyses data by using the Job ID, which was provided to the user at the first time of performing the simulation. To use this application, first the users are required to select the appropriate sub-section. If the user selects the first sub-section, a pathway list available in the BIOPYDB database will appear in the webpage. After selecting the name of the pathway of interest, the user will be redirected to the next page in which the users are required to provide initial values (binary values i.e. either 1 or 0) of the molecules (or nodes). If the initial values of the nodes are not provided in this section, then it will assign the random binary inputs (either 0 or 1) to the uninitialized nodes as defaults. The simulation engine running in the backend of BIOPYDB will instantly provide the Boolean equations of the intermediate molecules (i.e. the nodes/molecules which has an upstream regulator in the pathway network). It should be noted that

the Boolean equations generated in this section for the pathway species may not reflect the exact biological scenario as it is computationally generated from the “Interaction” table by using the information provided in the binary interactions of the pathway of interest. User can easily add new nodes and equations or modify the Boolean equation of the existing node in this interactive web interface. After confirming the initial values and the logical or Boolean equations, users are required to provide the total time steps (not more than 500) up to which the Boolean simulation will be performed. If the user does not provide this field then a default value (time steps 30) will be assigned to the submitted simulation. The user can also select the nodes of interest for which the time variation would be analyzed. After completing these steps, user can now submit the job in BIOPYDB web server. The logical simulation is performed with the help of popular python package BooleanNet [54]. A unique job ID will be provided to the user after submission of the simulation and the outputs will be provided in the next page. The nature of the stability (stable or cyclic attractor) of the simulation outcomes will be provided including the other parameters such as total transition states, number of input nodes, cycle length (if any), attractor length etc. The state transition data can also be downloaded as well as viewed in the web page. Temporal dynamics of the nodes of interest selected in the previous section will also be shown in the image and can be downloaded.

Similar analyses can also be performed on a new (or user-defined) pathway using the second sub-section of this application. Unique job ID will also be provided in this section, which can be viewed later by providing the job ID in the third sub-section of this application. The results of a performed simulation will be stored in the database for the next six months and until then it can be viewed by this application interface.

4.5.6 Dynamic Analyses

Dynamic analyses of the continuous temporal variations of the pathway species under a specific biological condition could be performed through this application. This application is based on the Ordinary Differential Equations (ODEs) analysis of the reaction/enzyme kinetics of the biochemical reaction cascades [48], [55]. This application can be performed on the existing pathway of BIOPYDB or on a new pathway model provided by the user. In order to perform the dynamic analyses of the existing pathway, users are required to select the pathway name from the drop-down list. After that, users will be asked to select the variables (molecular species) and its corresponding initial concentration in the HTML format used in the application interface. Simultaneously, users are required to provide the reaction kinetic equation (functional forms) of each selected variables/species of the pathway, which should be based on its functional relationships or reaction orders with the other interacting species. Users can also delete or insert any new molecule in the variable list and can execute the dynamic simulation of any number of variables. The users are also required to provide the numerical values of the rate parameters of the kinetic equations used for modeling the pathway reactions. Moreover, the users will be asked to provide the initial and final time (in seconds) up to which the simulation will be run and the time step at which the data for the time series will be saved. After submitting all these information, the user can submit the dynamic model into the BIOPYDB simulation engine by using the “Submit” button. The simulation outputs will show the time series plot of all the pathway components/variables within the time span provided by the user. On the other hand, to execute the similar dynamic analysis on a completely new pathway model, the user can select the “Ordinary Differential Equation of a new pathway” tab and then follow the same procedures described above. Depending on the size of the model (i.e. the number of independent variables and the complexities of the equations), the total time required for executing the entire simulation may vary. The fourth order Runge-Kutta numerical integration algorithm (implemented through C/C++ programming language) is used here for a quick simulation of the dynamic models [55].

4.5.7 Find Interaction

Using this user-friendly application, the user can search either all the interacting proteins of a particular protein of interest or the interaction between two specific proteins from BIOPYDB interactions data sets. In each case, it will give the interactions of the protein(s) as a directed network including a table with the appropriate PubMed and iRefIndex links of each queried interaction. Users can then directly check the literature reference of that queried interaction.

4.5.8 Find Disease

This application is useful for finding the diseases associated with a particular protein in the signaling network. This tool is divided into three parts, one for searching mutated proteins associated with a disease, second for searching diseases which are associated with a particular protein of interest and another is for searching the connection between a particular protein and disease from the protein-disease database: Malacards. BIOPYDB provides a beneficial interface for accessing, retrieving and viewing the protein-disease mapping data from this application. The Malacards database accession IDs are provided for each query performed through this tool.

4.5.9 Data Download Application

In order to facilitate the data downloading process more efficiently from the database, BIOPYDB has developed a useful and unique application written in PHP. It is unique because each time, upon user's request, this application initially generates the downloadable data (pathway image in SVG, PNG, and JPEG format or interaction data in Tab delimited, SBML or BIOPAX formats) from the raw data stored in the database and then serves for further downloading process. It allows the users to obtain an up to date data from the database and thus the developers of the database do not have to bother about the post-processing of the pathway data every time for downloading. This automated process also makes the database maintenance much easier and less time-consuming.

To make the database more accessible through various platforms (i.e. browsers, command line or any other third-party software), RESTful API service is also provided here. Under the "Download" section, elaborate descriptions of the protocols required to use the API service is mentioned. Through this service, users can access and obtain various pathway related data through browser or command line programming. It can provide the list of pathways with the BIOPYDB accession ID numbers; pathway images in SVG, PNG, and JPEG-encoded formats; pathway description and statistics encoded in XML; and information of pathway species, reactions, and diseases encoded in JSON formats. A list of URLs is available for obtaining such information through API in the database webpage under this section.

4.5.10 Web Interface

The web interface of BIOPYDB is designed in such a way so that users can easily interact and navigate through the database web pages. The "Homepage" of the database contains all the necessary tabs like "About BIOPYDB", "Browse pathways", "Downloads", "Tools", "Upload Pathway", "FAQ" and "Contact Us". The database homepage also provides text "Search" and "Advanced search" options. "Search" option has "auto-suggestion" option enabled, which facilitates more convenient searching of proteins, pathways, and diseases. The left panel of homepage contains "News & Updates" to highlight the subsequent updates and the right panel contains "Current Database Statistics" which is automatically updated if there is any changes/insertion or deletion in the database. In Figure 6, an HTML web page dumped picture is provided to understand the navigation of the web pages from the BIOPYDB home page.

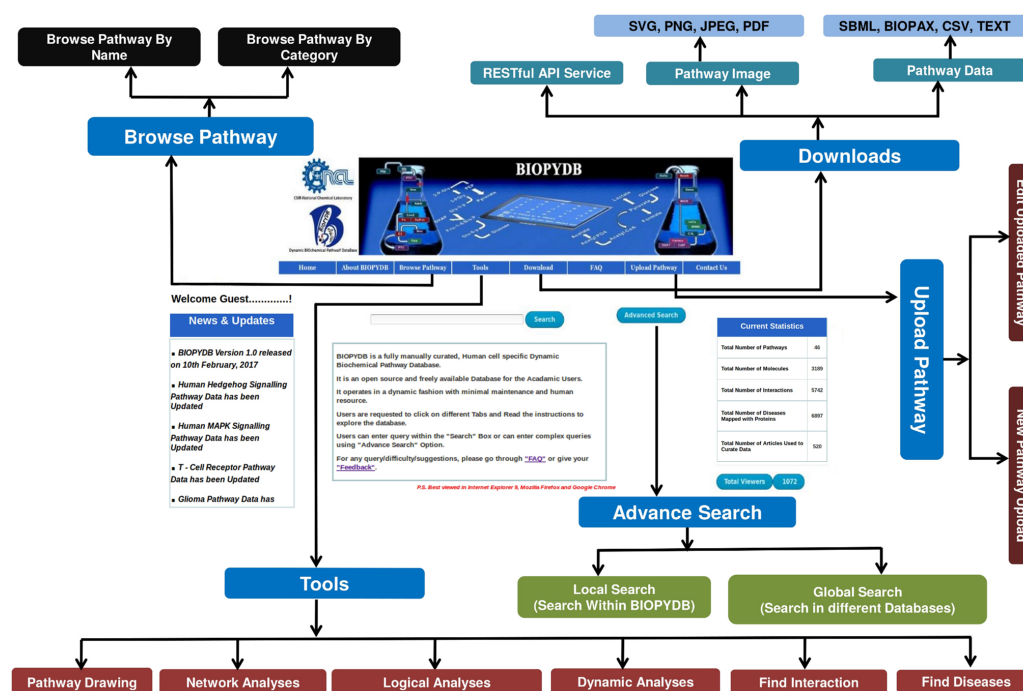


Figure 6: The web-interface of BIOPYDB.

5 Discussion

It is worth mentioning that BIOPYDB has a unique repertoire of Cell growth regulation, Developmental and Immune Signaling pathways, which gives a comprehensive understanding of the signaling pathways and their important roles in regulating the human cells/tissues. These pathways are very much useful to understand the de-regulations involved in diseases related to abnormal cell proliferation, immunity, birth defects and various others pathologic conditions. The exhaustive list of Interleukin pathways helps the users of BIOPYDB to gain a detailed insight into the signaling mechanisms of the immune-regulatory network that plays a pivotal role in understanding both infectious diseases and cancer. All the pathways are manually curated and possess up to date data synced with the current experimental findings. Cross validation of the pathway data is also available on the same platform with the other similar databases. It uses a new ontology tree, based on the pathway functions, and successively categorized the pathways accordingly. A unique pathway nomenclature system is introduced in this database to eliminate the ambiguities found in the existing databases [14]. In terms of the quality of data, the developers of BIOPYDB have rigorously cross-checked the pathway data (properly supported with existing and up-to-date literature) before including it in the main database. More number of targets or output proteins/molecules have been included in each of the pathways by deep literature mining, so that a comprehensive picture of the target or output products of biochemical reaction cascades could be shown and linked with various phenotypic outcomes of a cell or tissue. The intermediate molecules are also included in more number so that cross-talks with other pathway molecules could be easily and accurately depicted in the pathway diagram. For example, to show the Notch signaling events, its cross-talks with other important pathway proteins, such as JAK/STAT, HIF1, and P53 are also included. The inclusion of this information is required as these cross-talk molecules are playing major roles in regulating this pathway and hence it would be incomprehensive if these proteins are not included in the Notch signaling pathway data. The definition and the boundary of the reconstructed pathway diagram are not restricted to the core molecular entity of the newly reconstructed pathway, rather it is expanded as much as possible by including more number of the cross-talks through deep literature mining to provide a comprehensive knowledge of the pathway reactions. Due to this reason, the number of molecules of a particular pathway shown in BIOPYDB is higher than any of the similar pathway databases.

Another strong merit of BIOPYDB is that it not only provides the biochemical pathway data, but also provides a common platform to draw new pathway image as well as analyse Topological properties (such as Shortest Path between two proteins, Distribution of shortest paths, In-Degree, Out-degree, Total Degree and their distributions, Closeness, Betweenness and Eigenvector Centrality), Logical analysis (using discrete, semi-dynamic Boolean equations) and Dynamic analysis (using Ordinary Differential Equations) of the existing as

well as new, user defined pathways [46], [47], [48], [56], [57]. Moreover, the availability of Protein-Protein and Protein-Disease data (which are displayed as network view) provides an added benefit to this platform, which is not available in any other databases.

Several novel features, such as, automatic annotation or mapping of pathway molecules to other databases, network representation of protein-disease network, advanced data searching across 17 similar pathway and drug databases, user friendly pathway upload and update system for users, instant view of uploaded data by the users, dynamic pathway image drawing, runtime generation of pathway data for download, pathway data analyses platform to perform structural and dynamic analyses, sharing of pathway data in computer readable file formats (SBML and BIOPAX), and its potential usefulness in academic research, experimentalists, computational and theoretical biologists, etc. separates out BIOPYDB from other existing databases. The database is also made accessible through RESTful API service, which also helps the other third party software and database developers to use and analyze the BIOPYDB data.

6 Conclusion

BIOPYDB is a useful resource of biological data of human biochemical pathways; annotated proteins/genes that are involved in the pathways; protein-protein interactions, and protein-disease mapping. All the available data is manually curated from several literature resources and is available free of cost through the web browser and RESTful API services. BIOPYDB is a resource or repertoire of biochemical pathway data and also provides a method or strategy to make similar resources or a multi-functional platform, which would be easy to update and maintain, user-friendly, dynamic. It provides a single computational platform that would be useful for multiple tasks from data searching to data analyses. In short, BIOPYDB is not only a database, rather it is an information system, integrated with its own manually curated biochemical pathway database, a computational platform for accessing pathway data from the external sources, and providing web-services for *in-silico* pathway data modeling and simulation analyses. The data curation process of BIOPYDB is steadily increasing and in near future, it also promises to add more pathway information by including other types of pathways (such as metabolic, gene regulations etc.). The long-term goal of this database is to develop a common pathway data searching and computational analyses platform for performing biochemical pathway based modeling and simulations.

Availability and Requirements

BIOPYDB can be easily accessible through: <http://biopydb.ncl.res.in/biopydb/index.php> using any modern web browser such as IE, Mozilla, and Google Chrome, etc.

Acknowledgment

Saikat Chowdhury and Sutanu Nandi acknowledge the research fellowship provided by DST-INSPIRE fellowship program. Piyali Ganguli acknowledges the Junior Research Fellowship provided by CSIR-HRDG. The authors are also grateful to The Director, CSIR-National Chemical Laboratory, Pune, for providing the required infrastructures, computer server rooms etc. for the database.

Funding

Dr. Ram Rup Sarkar is supported by grants from the Council of Scientific and Industrial Research, XII Five Year Plan Project "GENESIS" (Funder Id: 10.13039/501100001412) (BSC0121), "HOPE" (BSC0114) and the project from SERB, Ministry of Science and Technology, Government of India (Funder Id: 10.13039/501100001843, No. EMR/2016/000516 Dated 19-12-2016).

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

- [1] Kitano H. Systems biology: a brief overview. *Science*. 2002;295:1662–4.
- [2] Fiehn O. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp Funct Genomics*. 2001;2:155–68.
- [3] Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L. Metabolite profiling: from diagnostics to systems biology. *Nat Rev Mol Cell Biol*. 2004;5:763–9.
- [4] Sarkar R. The Big Data Deluge in Biology: challenges and solutions. *J Inform Data Min*. 2016;1:14.
- [5] Bauer-Mehren A, Furlong LI, Sanz F. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol Syst Biol*. 2009;5:290.
- [6] Altmäe S, Esteban FJ, Stavreus-Evers A, Simón C, Giudice L, Lessey BA, et al. Guidelines for the design, analysis and interpretation of ‘omics’ data: focus on human endometrium. *Hum Reprod Update*. 2014;20:12–28.
- [7] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37:1–13.
- [8] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23:2507–17.
- [9] Ghosh S, Matsuoka Y, Asai Y, Hsin KY, Kitano H. Software for systems biology: from tools to integrated platforms. *Nat Rev Genet*. 2011;12:821–32.
- [10] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 1999;27:29–34.
- [11] Joshi-Tope G, Gillespie M, Vastrik I, D’Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*. 2005;33(Suppl 1):D428–32.
- [12] Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, Venugopal AK, et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol*. 2010;11:R3.
- [13] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. *Nucleic Acids Res*. 2009;37(Suppl 1):D674–9.
- [14] Chowdhury S, Sarkar RR. Comparison of human cell signaling pathway databases – evolution, drawbacks and challenges. *Database [Internet]*. 2015 [cited 2014 Dec 18]. Available from: <https://academic.oup.com/database/article/doi/10.1093/database/bau126/2433126>. DOI: 10.1093/database/bau126.
- [15] Petri V, Jayaraman P, Tutaj M, Hayman GT, Smith JR, De Pons J, et al. The pathway ontology – updates and applications. *J Biomed Semantics*. 2014;5:7.
- [16] Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Res*. 2002;30:42–6.
- [17] Pico AR, Kelder T, Van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. *PLoS Biol*. 2008;6:e184.
- [18] Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 2012;40:D742–53.
- [19] McDowall MD, Scott MS, Barton GJ. PIPs: human protein–protein interaction prediction database. *Nucleic Acids Res*. 2009;37:D651–6.
- [20] Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res*. 2016;45:D362–D368.
- [21] Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Curr Prot Bioinformatics*. 2016;54:1.30.1–1.30.33.
- [22] Liu X, Yu X, Zack DJ, Zhu H, Qian J. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*. 2008;9:271.
- [23] Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*. 2008;9:405.
- [24] Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. 2006;34:D535–9.
- [25] Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*. 2012;40:D841–6.
- [26] Prasad TK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database – 2009 update. *Nucleic Acids Res*. 2009;37:D767–72.
- [27] Rappaport N, Twik M, Plaschkes I, Nudel R, Stein TI, Levitt J, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res*. 2016;45:D877–87.
- [28] Ellson J, Gansner E, Koutsofios L, et al., editors. *Graphviz – open source graph drawing tools*. Graph Drawing. Berlin, Heidelberg: Springer; 2001. p. 483–4.
- [29] Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, et al. The BioPAX community standard for pathway data sharing. *Nat Biotechnol*. 2010;28:935–42.
- [30] Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*. 2003;19:524–31.
- [31] Webb RL, Ma’ayan A. Sig2BioPAX: Java tool for converting flat files to BioPAX Level 3 format. *Source Code Biol Med*. 2011;6:5.
- [32] Frank MR, Omiecinski ER, Navathe SB, editors. *Adaptive and automated index selection in RDBMS*. Advances in Database Technology – EDBT ’92. EDBT. Berlin, Heidelberg: Springer; 1992. p. 277–92.
- [33] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res*. 2013;41:D991–5.
- [34] Consortium U. Activities at the universal protein resource (UniProt). *Nucleic Acids Res*. 2014;42:D191–8.
- [35] Mann RK, Beachy PA. Cholesterol modification of proteins. *Biochim Biophys Acta Mol Cell Biol Lipids*. 2000;1529:188–202.
- [36] Cohen P. Review lecture: protein phosphorylation and hormone action. *Proc R Soc Lond B Biol Sci*. 1988;234:115–44.

- [37] Willard FS, Crouch MF. Nuclear and cytoskeletal translocation and localization of heterotrimeric G-proteins. *Immunol Cell Biol.* 2000;78:387–94.
- [38] Gaston K, Jayaraman P-S. Transcriptional repression in eukaryotes: repressors and repression mechanisms. *Cell Mol Life Sci.* 2003;60:721–41.
- [39] Pribluda VS, Pribluda C, Metzger H. Transphosphorylation as the mechanism by which the high-affinity receptor for IgE is phosphorylated upon aggregation. *Proc Natl Acad Sci USA.* 1994;91:11246–50.
- [40] Pawson T, Scott JD. Signaling through scaffold, anchoring, and adaptor proteins. *Science.* 1997;278:2075–80.
- [41] Crabtree GR. Calcium, calcineurin, and the control of transcription. *J Biol Chem.* 2001;276:2313–6.
- [42] Komander D. The emerging complexity of protein ubiquitination. *Biochem Soc Trans.* 2009;37(Pt 5):937–53.
- [43] Mullor JL, Sánchez P. Pathways and consequences: Hedgehog signaling in human disease. *Trends Cell Biol.* 2002;12:562–9.
- [44] Nusse R. Wnt signaling in disease and in development. *Cell Res.* 2005;15:28–32.
- [45] Polakis P. Wnt signaling and cancer. *Genes Dev.* 2000;14:1837–51.
- [46] Chowdhury S, Pradhan RN, Sarkar RR. Structural and logical analysis of a comprehensive hedgehog signaling pathway to identify alternative drug targets for glioma, colon and pancreatic cancer. *PLoS One.* 2013;8:e69132.
- [47] Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 2004;5:101–13.
- [48] Lee JM, Gianchandani EP, Eddy JA, Papin JA. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput Biol.* 2008;4:e1000086.
- [49] Csardi G, Nepusz T. The igraph software package for complex network research. *Int J Complex Syst.* 2006;1695:1–9.
- [50] Singh A, Nascimento JM, Kowar S, Busch H, Boerries M. Boolean approach to signalling pathway modelling in HGF-induced keratinocyte migration. *Bioinformatics.* 2012;28:i495–501.
- [51] Fumiã HF, Martins ML. Boolean network model for cancer pathways: predicting carcinogenesis and targeted therapy outcomes. *PLoS One.* 2013;8:e69008.
- [52] Müssel C, Hopfensitz M, Kestler HA. BoolNet – an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics.* 2010;26:1378–80.
- [53] Saez-Rodriguez J, Simeoni L, Lindquist JA, Hemenway R, Bommhardt U, Arndt B, et al. A logical model provides insights into T cell receptor signaling. *PLoS Comput Biol.* 2007;3:e163.
- [54] Zhang R, Shah MV, Yang J, Nyland SB, Liu X, Yun JK, et al. Network model of survival signaling in large granular lymphocyte leukemia. *Proc Natl Acad Sci USA.* 2008;105:16308–13.
- [55] Butcher JC. The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods. New York: John Wiley & Sons Inc, 2008.
- [56] Chowdhury S, Sarkar R. Drug targets and biomarker identification from computational study of human notch signaling pathway. *Clin Exp Pharmacol.* 2013;3:2161–1459.1000137.
- [57] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol.* 2012;8:e1002375.

Supplemental Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/jib-2017-0072>).