Chapter 1

Hierarchical Biological Pathway Data Integration and Mining

Shubhalaxmi Kher, Jianling Peng, Eve Syrkin Wurtele and Julie Dickerson

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/49974

1. Introduction

Biological pathway data is the key resource for biologists worldwide. Interestingly, most of these sources that generate, update, and analyze data are open source. One of the observations that motivated this research work is that, the repositories of data created by a variety of laboratories and research units worldwide represent same pathways with significant details. Generally, if the pathway data has resulted from experimentation, then it is expected that across different resources, under similar conditions, pathways would be exactly identical and biologists may pickup from any source. Interestingly, almost all of the biological data sources refer to data integration of some kind. It may involve rigorous integration mechanisms within the data source and the purpose of integration may change the perspective of looking at the integration.

These efforts in integration may be either local to the source or lack details associated with integration within a pathway, across pathways, or from various data sources etc. Further, the key attributes or design criteria may not be well documented and or may not be readily available to the biologist. In other words, the integration may be achieved as vertical integration (within the data source), or horizontal integration (across data sources). Since most of the extensively integrated data sources (plants or humans) like BioCyc-level-I, Reactome are human curated, it is hard to identify the integration done by the sources like; BioCyc. Also, on a similar note, it may not be apparent to find exactly when the data was integrated looking at a pathway.

Data in general refers to a collection of results, including the results of experience, observation, or experiment, or a set of premises and can be utilized at the maximum when made available to all in a common format. Different organizations and research laboratories around the world store the data in their own formats; this diversity of data sources is caused due to many factors including lack of coordination among the organizations and research

laboratories. These intellectual gaps can be bridged by adopting new technology, mergers, acquisitions, and geographic coordination of collaborating groups [1].

For the open source biological databases, it is common for the biologists and researchers to refer to many databases in order to pursue inference or analysis; though it is one of the most challenging tasks. Biological pathway data integration is aimed to work with repositories of data from a variety of sources. As such, two or more databases may not provide identical information for a given pathway, but integrating these two databases may yield a richer resource for analysis. Additionally, the conditions under which data is collected, either by experimentation or by collecting evidence of the published material, in either case the supporting references play a crucial role and is of interest to the biologists in making the analysis more meaningful. At present there are over 200 biological pathway databases. However, very few of them are independently created. Some of these databases may be derived from different data sources. Unfortunately, the documentation often does not reveal details of the data collection, sources, and dates. Further, the research groups involved in analysis of the data usually selectively use data from a single data source. For example, for yeast studies, the Saccharomyces Genome Database (SGD) is the reference for most analyses [2].

In case of biological pathway data, rapid accumulation of genomic and proteomic data have made two major bioinformatics problems apparent.

- The lack of communication between different bioinformatics data resources; whether they are databases or individual analysis programs.
- Biological data are hierarchical and highly related yet are conventionally stored separately in individual database and in different formats.
- Additionally, they are governed more by how data is obtained rather than by what they mean.

Most commercially available bioinformatics systems perform functional analysis using a single data source; an approach that emphasizes pathway mapping and relationship inference based on the data acquired from multiple data sources. Each pathway modality in the data has its own specific representation issues which must be understood before attempting to integrate across modalities.

1.1. Overview

There has been a dramatic increase in the number of large scale comprehensive biological databases that provide useful resources to the community like; Biochemical Pathways (KEGG, AraCyc, and MapMan), Protein Interactions (biomolecular interaction network database), or systems like; Dragon Plant Biology Explorer and Pathway Miner for integrating associations in metabolic networks and ontologies [3-8]. Other databases such as Regulon DB, PlantCARE, PLACE, EDP:Eurokaryotic promoter database, Transcription Regulatory Regions Database, Athamap, and TRANSFAC store information related to transcriptional regulation[9-15].

The aim of molecular biology is to understand the regulation of protein synthesis and its reactions to external and internal signals. All the cells in an organism carry the same genomic data, yet their protein makeup can be drastically different; both temporally and spatially, due to regulation. Protein synthesis is regulated by many mechanisms at its different stages. These include mechanisms for controlling transcription initiation, RNA splicing, mRNA transport, translation initiation, post-translational modifications, and degradation of mRNA/protein. One of the main junctions at which regulation occurs is mRNA transcription. A major role in this machinery is played by proteins themselves that bind to regulatory regions along the DNA, greatly affecting the transcription of the genes they regulate [16]. Friedman introduces a new approach for analyzing gene expression patterns that uncovers properties of the transcriptional program by examining statistical properties of dependence and conditional independence in the data.

For protein interactions, it is intended to connect related proteins and link biological functions in the context of larger cellular processes [17]. The content of these data sources typically complements the experimentally determined protein interactions with the ones that are predicted from gene proximity, fusion, co-expressed data, as well as those determined by using phylogenetic profiling. Each pathway modality in the data has its own specific representation issues which must be understood before integration across modalities is attempted. At present, the bioinformatics database owner only develops private system to provide user with data query and analysis services; such as NCBI develops Entrez database query system which is used on GenBank. European Molecular Biology Laboratory (EMBL) develops Sequence Retrieval Systems. The EMBL Nucleotide Sequence Database maintained at the European Bioinformatics Institute (EBI), incorporates, organizes, and distributes nucleotide sequences from public sources [18]. The database is a part of an international collaboration with DDBJ (Japan) and GenBank (USA). Data are exchanged between the collaborating databases on a daily basis to achieve optimal synchrony. The key point is how to share the heterogeneous databases and make a common query platform for users [19].

Friedman [16] describes early microarray experiments that examined few samples and mainly focused on differential display across tissues or conditions of interest. Such experiments collect enormous amounts of data, which clearly reflects many aspects of the underlying biological processes. An important challenge is to develop methodologies that are both statistically sound and computationally tractable for analyzing such data sets and inferring biological interactions from them. Most of the analysis tools currently used are based on clustering algorithms. The clustering algorithms attempt to locate groups of genes that have similar expression patterns over a set of experiments. Such analysis has proven to be useful in discovering genes that are co-regulated and/or have similar function. A more ambitious goal for analysis is to reveal the structure of the transcriptional regulation process. This is clearly a hard problem. Not only the current data is extremely noisy, but, mRNA expression data alone only gives a partial picture that does not reflect key events such as; translation and protein (in) activation. Finally, the amount of samples, even in the largest experiments in the foreseeable future, does not provide enough information to construct a fully detailed model with high statistical significance.

Some conventional bioinformatics approaches identify hypothetical interactions between proteins based on their three dimensional structures or by applying text mining techniques. Emerging protein chip technologies are expected to permit the large scale measurement of protein expression levels. Corresponding structural data are stored in data source such as protein data bank and represent invaluable sources of understanding of protein structures, functions and interactions. Successful use of high throughput protein interaction determination techniques such as yeast two hybrids, affinity purification followed by mass spectrometry and phage display has shifted research focus from a single gene/protein to more coherent network perspectives. Large scale protein-protein interaction data and their complexes are currently available for a number of organisms and data are stored in several interaction data sources such as BIND [6], DIP [20], IntAct [21], GRID [22] and MINT [23] that is all equipped with basic bioinformatics tools for protein network analysis and visualization. INCLUSive is a web portal and service registry for microarray and regulatory sequence analysis [24]. This provides a comprehensive index for all data integration research projects.

The integration and management technique of heterogeneous sequence data from public sequence data source is widely used to manage diverse information and prediction. It is important for the biologists to investigate these heterogeneous sources and connect the public biological data source and retrieve sequences which are similar to sequences they have, and the results of their retrieval are used in homology research, functional analysis, and predication. However, there are few software packages available to deal with the sequence data in most biological laboratories and they are stored in file formats. File formats is another important issue for biological pathway data sources. XMl, SBML (systems biology markup language), KBML (KEGG), BSML (Bioinformatic Sequence Markup Language) based on XML, and a variety of versions of XML are used for representing the complex and hierarchical biological data. Each flat file from public biological database has different format. Recent tools which convert formats among standards are implemented in JAVA or Perl module. The constraints associated with biological pathway formats are the following;

- Conversion among different formats needs different parsers to extract the user interesting field.
- Formats can be modified anytime.
- Understand the range of field, its value is difficult, and data types in the same field in each format can be different.

From the discussions above, one of the major challenges of the modern bioinformatics research is therefore to store, process, and integrate biological data to understand the inner working of the cell defined by complex interaction networks. Additionally, the integration mechanisms may not register the important details like, copies of inputs files and time of integration along with the integrated output file.

In this chapter, issues related to biological pathway data integration system are discussed and a user friendly data integration algorithm across data sources for biological pathway, particularly, metabolic pathway as a case is presented. i.e. the data integration (BPDI) algorithm that integrates pathway information across data sources and also extracts the

abstract information embedded within them are addressed. Today, a bioinformatics information system typically deals with large data sets reaching a total volume of about one terabyte [25]. Such a system serves many purposes;

- User can select the data sources and assign confidence to each selected data source
- It organizes existing data to facilitate complex queries
- It infers relationships based on the stored data and subsequently predicts missing attribute values and incoming information based on multidimensional data.
- Data marts (extension of data warehouse) support different query requests.

2. Data management and integration

The Pathway Resource List contains over 150 biological pathway databases and is growing [26]. Usually, first step for the user is to identify a subset of these data sources for integration. To consolidate all the knowledge for a particular organism, extract the pathways from each database need to be extracted and transformed into a standard data representation before integration. Representation of the pathway data in each data source poses another challenge as each pathway modality has its own specific representation issues which must be understood before attempting integration across modalities. For example, metabolic pathways, signal transduction pathways, protein-protein interaction, gene regulation etc.

Commonly employed styles of data integration may be implemented in different contexts and under requirements, in order to reuse the data across applications for research collaboration. Some of the data integration and management efforts are presented in [27-32]. Several major approaches have been proposed for data integration, which can be roughly classified into five groups [33-34] namely; data warehousing, federated databasing, service-oriented integration, semantic integration and wiki-based integration. Across all of these groups, to a significant extent, an increasingly important component of data integration is the community effort in developing a variety of biomedical ontologies to deal in a more specific manner with the technicality and globality of descriptors and identifiers of information that has to be shared and integrated across various resources. Variety of approaches for data integration is discussed below.

Data Warehousing

The data warehouse approach offers a “one-stop shop” solution to ease access and management of a large variety of biological data from different data sources. The user does not need to access many web sites for multiple data sources. Despite its advantages, the data warehouse approach has a major problem; it requires continuous and often human-guided updates to keep the data comprehensive of the evolution of data sources, resulting in high costs for maintenance. Many biological data sources change their data structures roughly twice a year.

Data integration with Federated Approach

Unlike data warehousing (with its focus on data translation), federated databasing focuses on query translation. The federated database fetches the data from the disparate data

sources and then displays the fetched data for its user base. Queries in federated databases are executed within remote data sources and results displayed in federated databases are extracted remotely from the data sources. Due to this capability, federated databasing has two major advantages.

- Federated databases can be regarded as an on-demand approach to provide immediate access to up-to-date data deposited in multiple data sources.
- Compared with data warehousing, federated databasing does not replicate data in data sources; therefore, it presents relatively inexpensive costs for storage and curation. However, federated databasing still has to update its query translation to keep pace with data access methods at diverse remote data sources.

Service –Oriented Approach

A decentralized approach is also being developed, in which individual data sources agree to open their data via Web Services (WS). The service-oriented approach enables data integration from multiple heterogeneous data sources through computer interoperability. The service-oriented approach features data integration through computer-to-computer communication via Web API and up-to-date data retrieval from diverse data sources. Heterogeneous data integration requires that many data sources should become service providers by opening their data via WS and by standardizing data identities and nomenclature to ease data exchange and analysis.

Semantic Web

Most web pages in biological data sources are designed for human reading. RDF provides standard formats for data interchange and describes data as a simple statement, containing a set of triples: a subject, a predicate, and an object. Any two statements can be linked by an identical subject or object. OWL builds on RDF and Uniform Resource Identifier (URI) and describes data structure and meaning based on ontology, which enables automated data reasoning and inferences by computers. Application of semantic Web technologies is a significant advancement for bioinformatics, enabling automated data processing and reasoning. The semantic integration uses ontologies for data description and thus represents ontology-based integration. [27] reviews the current development of semantic network technologies and their applications to the integration of genomic and proteomic data. His work elaborates on applying a semantic network approach to modeling complex cell signaling pathways and simulating the cause-effect of molecular interactions in human macrophages. [31] Illustrates his approach by comparing federated approach versus warehousing versus semantic web using multiple sources.

Wiki-based Integration

A weakness common to all the above approaches is that the quantity of users' participations in the process is inadequate. With the increasing volume of biological data, data integration inevitably will require a large number of users' participations. A successful example that harnesses collective intelligence for data aggregation and knowledge collection is Wikipedia: an online encyclopedia that allows any user to create and edit content. It is

infeasible to integrate such large amounts of data into a single point (such as a data warehouse). Data sources are developed for different purposes and fulfill different functions. Therefore, it is promising to establish an efficient way for data exchange among these distributed and heterogeneous data sources. However, a dozen of data sources are designed merely for data storage, but not for data exchange.

2.1. Survey of Pathway Databases and Integration Efforts

Table 1 below shows various data integration efforts and projects for biological pathways worldwide.

Biochemical pathways	Description
BRITE	Bio molecular Relations in Information Transmission and Expression
EcoCyc/MetaCyc	Encyclopaedia of E. coli genes and metabolism; Metabolic encyclopedia
EMP	Metabolic pathways
KEGG	Kyoto encyclopaedia of genes and genomes
Biochemical Pathways	Enzyme database and link to biochemical pathway map
Interactive Fly	Biochemical pathways in Drosophila
Metabolic Pathway	Metabolic pathways of biochemistry
Molecular interaction	Kohn molecular interaction maps
Malaria parasite	Malaria Parasite metabolic pathways
aMAZE	Protein function and biochemical pathways project at EBI
PathDB	Metabolic pathway information
UM-BBD	Microbial bio catalytic reactions and biodegradation pathways primarily for xenobiotic, chemical compounds
WIT	Function assignments to genes and the development of metabolic models
THCME Medical Biochemistry	Description of several metabolic and biochemical pathways
Signaling pathways	
Apoptosis	Pathways of apoptosis at KEGG
BBID	Database of images of biological pathways, macromolecular structures, gene families, and cellular relationships
BioCarta	Several signalling pathways
BIND	The bio molecular interaction network database

CSNDB	Cell signalling networks database
GeneNet	Information on gene networks, groups of co-ordinately working genes
GeNet	Information on functional organization of regulatory gene networks
SPAD	Signalling pathway database
STKE	Pathway information
TransPath	Pathways involved in the regulation of transcription factors
Protein-protein interactions	
Blue Print	Biological interaction database
CYGD	Protein-protein interaction map at Comprehensive Yeast Genome Database
CytoScape	Visualization and analysis of biological network
DIP	Database of interacting proteins
GenMAPP	Gene Map Annotator and Pathway Profiler
GRID	The General Repository for Interaction Datasets
Proteome Bio knowledge	Biological information about proteins comprise Incyte's Proteome Bio Knowledge Library
Protein Interaction Domains	Signal transduction
Reactome	A knowledgebase of biological processes
Yeast Interaction Pathway	PathCalling Yeast Interaction Database at Curagen

Table 1. Various Data integration Efforts

Other efforts towards designing new applications for data mining and integration at the K.U.Leuven Center for Computational Systems Biology include;

- aBandApart (2007): A software to mine MEDLINE abstracts to annotate human genome at the level of cytogenic bands.
- ReModiscovery (2006): An intuitive algorithm to correlate regulatory programs with regulators and corresponding motifs to a set of co-expressed genes
- LOOP (2007): A toll to analyze ArrayCGH loop designs. ArrayCGH is a microarray technology that can be used to detect aberrations in the ploidy of DNA segments in the genome of patients with congenital anomalies.
- SynTReN (2006): A generator of synthetic gene expression data for design and analysis of structure learning algorithms.
- BlockAligner (2005): Provides an API in R to query BioMart databases such as Ensemble.
- BlockSampler (2005): Finds conserved blocks in the upstream region of sets of orthologous genes.

- M@cBETH (2005) (a Microarray Classification Benchmarking Tool on a host server): Web service offers the microarray community a simple tool for making optimal two class predictions.
- TxTGate (2004): A literature index database designed towards the summarization and analysis of groups of genes based on text.
- Endeavour is a software application for the computational prioritization of test genes based on training genes using different information sources such as MEDLINE abstracts and LocusLink textual description, gene ontology, annotation, BIND protein interactions, and Transcription Factor Binding Sites (TFBS).
- TOUCAN2 (2004): A workbench for regulatory sequence analysis on metazoan genomes: Comparative genomics detection of significant transcription factor binding sites and detection of cis-regulatory modules in sets of coexpressed/ coregulated genes.
- INCLUSive (2003): A suit of algorithms and tools for the analysis of gene expression data and the directory of cis-regulatory sequence elements.
- Adaptive Quality-Based Clustering (AQBC) (2002): AQBC is a heuristic, iterative two-step algorithm to cluster gene expression data.
- MotifSampler (2001): Finds over represented motifs in the upstream region of a set of co-regulated genes.

2.2. Types of pathways

Biological networks are studied and modeled at different description levels establishing different pathway types, For example; metabolic pathways describe the conversion of metabolites by enzyme-catalyzed chemical reactions given by their stoichiometric equations, such as the main pathways of the energy household as Glycolysis or Pentose Phosphate pathway. Another pathway type is signal transduction pathways, also known as information metabolism, explaining how cells receive, process, and responds to information from the environment. A brief description about various types of pathways is given below.

A. Metabolic Pathways describe the network of enzyme-catalyzed reactions that release energy by breaking down nutrients (catabolism) and building up the essential compounds necessary for growth (anabolism). Experimentally determined metabolic pathways have established for a few model organisms, but most metabolic pathways databases contain pathway data that has been computationally inferred from the genomes annotations. Because most genome annotations are incomplete, metabolic pathway databases contain pathway holes which can only be addressed by experiment or computational inference. A good test of a reconstructed metabolic network is to ask if it can produce the set of essential compounds necessary for growth, given a known minimal nutrient set. To solve this problem, metabolism can be represented as a bipartite directed graph, where one set of nodes represents metabolites, the other set represents biochemical reactions with labeled edges used to indicate relationships between nodes (reaction X produces metabolite Y, or metabolite Y is-consumed-by reaction X.

B. Gene Regulatory Networks describe the network of transcription factors that bind regulatory regions of specific genes and activate or repress their transcription. Gene regulatory networks or transcription networks have been found to contain recurring biochemical wiring patterns, termed network motifs, which carry out key functions. How does one find the most significant recurring network motif in a given transcriptional network? To answer this question, transcription networks can be described as directed graphs, in which nodes are genes, and edges represent transcription interactions, where a transcription factor encoded by one gene modulates and transcription rate of the second gene.

C. Signaling Pathways describe biochemical reactions for information transmission and processing. Unlike metabolic pathways that catalyze small molecule reactions, signaling pathways involve the post translational modification of proteins leading to the downstream activation of transcriptional factors. They are often formed by cascades of activated/deactivated proteins or protein complexes. Such signal transduction cascades may be seen as molecular circuits which mediate the sensing and processing of stimuli. They detect, amplify and integrate diverse external signals to generate responses, such as changes in enzyme activity, gene expression, or ion channel activity. Integration of signaling pathways poses a greater challenge than with metabolic pathways because of diversity of representation schemes for signaling. Some Signaling databases like; PATIKA [35] and INHO [36] use compound graphs to represent signaling pathways, while other object oriented databases use inheritance to establish relationships between post translational modifications of proteins.

D. Protein-Protein Interaction: In proteomic analysis, target genes are used as bait in immuno-precipitation to identify potential binding patterns in cell lysate. The higher level databases such as; KEGG [3], TRANSPATH [37], ReactomeSTKE [38], and MetaCyc [39] networks of interacting proteins with definite cellular processes including metabolism, signal transduction and gene regulation. These resources typically represent biological information in the form of individual pathway diagrams summarizing experimental results collected during years of research on particular cellular functions. Currently, no single method is capable of predicting all possible protein interactions and such integrative resources as SPRING and predictome combine multiple theoretical approaches to increase prediction accuracy and coverage. A problem with these networks is the high number of false alarms.

E. Ontology Vocabulary Mapping: Ontology provides a formal written description of a specific set of concepts and their relationships in a particular domain. GO ontology has three categories molecular function, biological process and cellular composition. Integration of signaling pathways poses a greater challenge than with metabolic pathways because of the diversity of representation schemes for signaling.

2.3. Integration issues

Biological plant pathway data integration is a multi-step process. It includes integration of various types of pathways, interactions, and gene expression. On another level, it includes

various species and different databases. A hierarchical pathway data integration scheme is presented in Figures 1 and 2 below.

Each database also defines supporting evidence codes specifically defined to consider criteria for selection, however may not be explicitly illustrated and that may not be similar across various sources. This heterogeneity in evidence codes and their representation needs consideration [40]. Since the evidence code may originate as a result of experimentation or as evidence from published text, integration of the plant pathway data across databases involves standardizing the evidence code prior to the integration. The first step is to integrate the evidence codes for a given pathway across database. Biological databases are results of experiments carried out with different conditions and controls, mostly open source, and employs a variety of formats [41]. Integrating such databases is a multi-step procedure and involves handling the complexities associated with heterogeneous data integration.

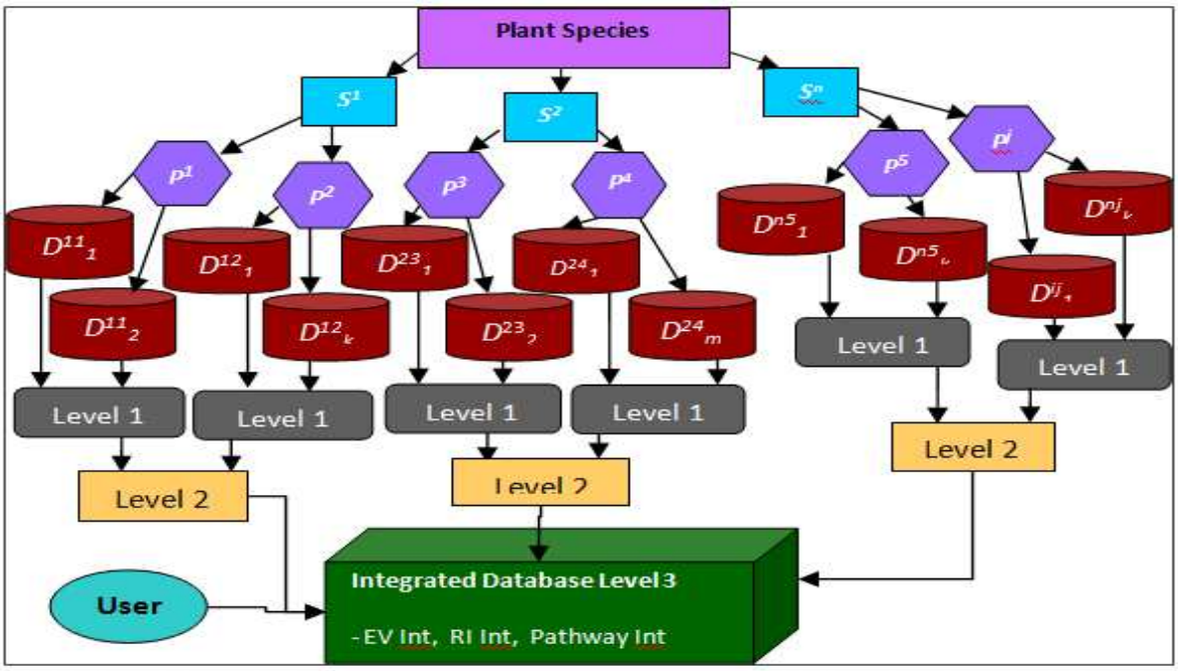


Figure 1. Hierarchical Pathway data Integration Scheme

A. Ontology Development

Since isolation of ontologies complicates data integration, so in order to use ontologies at their full potential, concepts, relations, and axioms must be shared when possible. Domain ontologies must also be anchored to an upper ontology in order to enable the sharing and reuse of knowledge.

B. Synonym Integration

While integrating information about a pathway from a database, entities require independent approach. One such entity is synonym. Each database lists a set of synonyms that need integration to configure a pool of synonyms without causing duplication. In the

data integration platform developed the synonym integration has issues like avoiding duplication and accommodating number of synonyms associated with one entity. Some pathways may include two compounds with different names but having same empirical formula. In such cases integration is challenging as biologists may be further interested in reviewing the chemical structure along with the integrated output. However, almost all biological pathways are vertically extendable and can associate further details. The point here is to include all the salient features (from a biologist's standpoint) of the pathway. There is no thumb rule to define biologist's interests.

C. Evidence Codes and issues

For defining an evidence code with an entity, granularity is another variable. Depending on the database, EV may be either for an entity within a pathway such as a gene, a compound, reaction or enzyme or for the pathway itself. In other words, many databases use the same evidence code for an entire pathway and map that code to each interaction in the pathway. Others assign different EV codes to each interaction and sometimes to each compound or gene.

The Gene Ontology (GO) defines a set of thirteen EVs that assign evidence to gene function. BioCyc defines a class hierarchy structure of four basic EVs with subclasses. MetNetDB incorporates four EVs [42]. KEGG defines only one EV. Ideally, the EVs also reflect on the individual nodes within a specific pathway. Figure 2 depicts the data integration platform highlighting multiple data sources and integration based on user inputs.

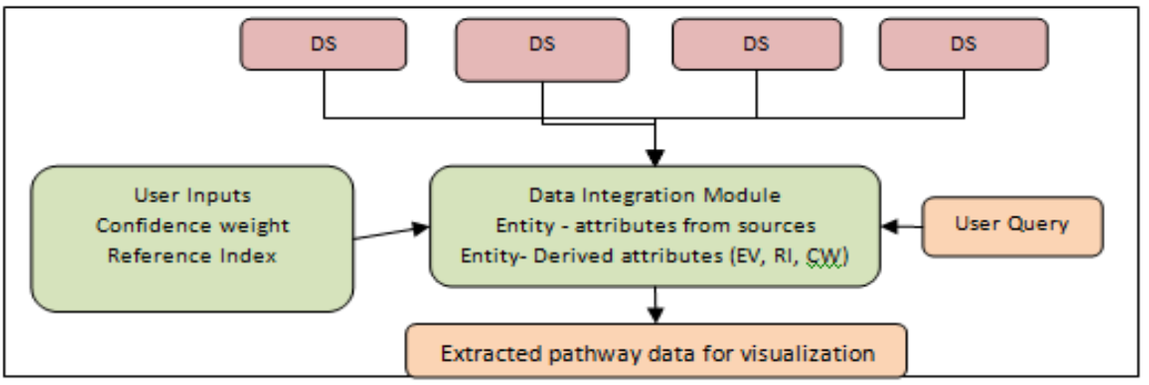


Figure 2. Data Integration Platform

1. Many databases use the same evidence code for an entire pathway and map that code to each interaction in the pathway. Others assign codes to each interaction and sometimes each compound or gene. In other words, the granularity to which we can assign an EV may be either an entity such as a gene, a compound, reaction or enzyme within or across the pathway itself. The Gene Ontology (GO) defines a set of thirteen EVs that assign evidence to gene function [43]. BioCyc defines a class hierarchy structure of four basic EVs with subclasses [17]. MetNetDB incorporates four EVs. KEGG defines only one EV. Ideally, the EVs also reflect on the individual nodes within a specific pathway.

2. Since pathway information cannot be assessed with any reliability, it is hard to assign a measure of the orrectness/authenticity to any one database. We propose assignment to be user selective to resolve the issue. To combine the information, a heuristic rule set computes the composite EVs for the integrated database. The unification can be done using any one EV code set as a key. Since each database follows their own standard, it is likely that EVs may not find a perfect match among the databases or that there may be more than one likely match. To handle these situations, two matching sets, a *perfect match* and a *likely match* are considered. The EVs to find a match for *IEP* and *ND* from GO in EV set above with those in BioCyc result in more than one likely match *{GO: IEP → BioCyc: EV1, BioCyc: EV2}.*
3. Integrated Evidence Code (EVint) for Perfect Matches: The EV codes encompass the quantitative information giving an insight into how the data was obtained. They define the conditions/ constraint associated with obtaining the data.
4. Computing the Reference Index (RIint)
   For biological databases, the pathway information is mostly inferred by the curators based on experimental, computational, literature or other evidence. The references associated with the database are mostly accounted as a measure of support for the data. We introduce a qualitative approach to associate the references supporting the pathway or organism (or compounds or reactions). The reference index *RIint* is computed using a heuristic:

   1. For *Rank = High*, Ignore *VF.*
   2. For Rank = Low, Use only VF.
   3. For all other combinations of *Rank* and *VF,* compute the average.

Citations may be a robust way of supporting the claim in a database. However, some journals are ranked over other journals and citations from those journals will be valued more than citations in other sources. To accommodate this, we associate ranks with the journals. The *Rank* specifies the order of importance of journal as designated by the user. Additionally, we classify citations based on both the journal *Rank* and the *value factor (VF)*. Finally, based on the *Rank* and *VF,* the *Reference index (RI)* is computed.

3. Evidence codes integration algorithm

Given: Set of *n* databases *{D1, D2, D3, D4,……, Dn},*

(For illustration, only three data sources namely, Bio-Cyc, KEGG and MetNetDB are considered)

User input: C*onfidence weight (CW)*

List: Evidence Codes (EVi) for the object/entity *(Ei)* among the databases *(Di),*

for example; *D1/E1 {EV1}*, *D2/E1 {EV2}*,….

The steps below list the mapping process.

Step 1. For a given pathway/organism/entity,

List: *EV codes across the databases*. (See Tables III(a) and III(b))

Assign: ***Direct*** *= 1.0;* ***Indirect*** *= 0.8;* ***Computational*** *=0.6;* ***Hypothetical*** *= 0.5.*

Step 2. EV Unification (Rule Set –I)

BioCyc is a collection of 371 pathway/genome databases. Each pathway/genome database in the BioCyc collection describes the genome and metabolic pathways of a single organism. It considers a class hierarchy with four main classes. Since BioCyc and MetNetDB virtually use the same number of EV codes, the mapping is framed considering four major EV codes. KEGG uses only one EV for pathways namely '*manually entered from published materials'*. The EV code for KEGG to *Direct* is mapped using the rules like;

If Di = BioCyc/AraCyc/MetaCyc, and EV = EV-Exp, then Change EV = Direct

Unification of the EV codes for the databases is based on the expert knowledge. EV code mapping is done with respect to a reference data source and unified according to the set of rules above.

Step 3. Confidence Weight (*CWi*) Assignment

Researchers typically have databases that they treat as favored sources for different types of information. Since there is no precise rule for deciding which database is more correct and up to date, a user defined score, a *confidence weight (CW)* is applied. The EV mapping process is interactive and provides flexibility in choice for databases. Confidence is defined as,

CWi = {Very Strong, Strong, Moderate, Poor, Very Poor}
For example: *CW* ***KEGG****: Strong*, *CW****BioCyc***: *Normal*

Step 4. EVint (Rule Set-II)

Using heuristic rules, integrated EVcode is calculated.

Step 5. Decode *EVint* value

The EV value from Step 4 is decoded using:
$EVint = \Sigma\ (CWi^{*}\ EV)/|i| = x$

Step 6. Rank Index

- Rank the journals in their order of importance.
- Make an ordered list of journals assigning *Rank.*
- Rank the conferences in order of their importance.
- Make an ordered list of conferences.
- *Assign:*

  If the publication in not in the list, Then, *Rank = low*
  Else, *Rank = as defined by the list*

Step 7. Value Factor *(VF)*

The *VF* measures support for the entity using the publication evidence. This is a quantitative index with a temporal function.

For *t* = current year, compute *VF (t)* = |*P (t-2)*| / |*P*| where,

|*P (t-2)*| = Number of publications in the last *(t-2)* years for *Di*, and
|*P*| = Total number of publications listed in *Di*.

Step 8. Reference Index *(RIint)*

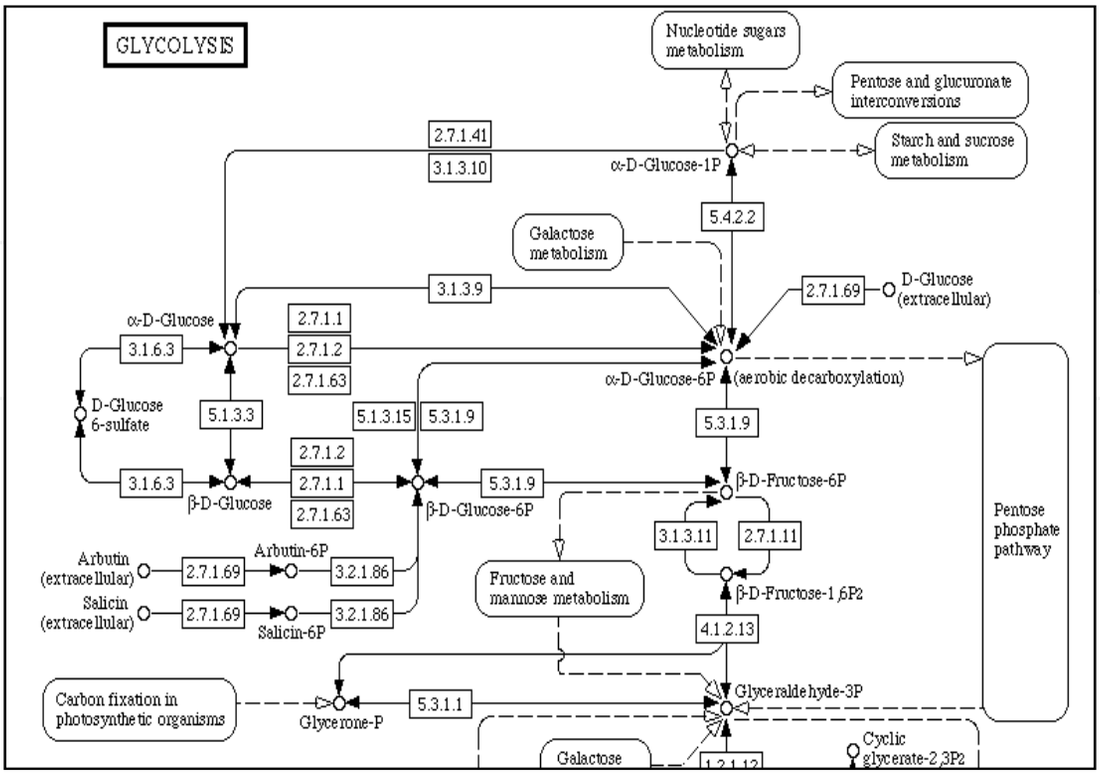
- Compute *RI* for *{D1,…Dn}* given by,
  RIi{t} = f {Rank, VF}
- Compute *RIint for a* pathway as;
  RIint = max {RIi}

3.1. Integration models

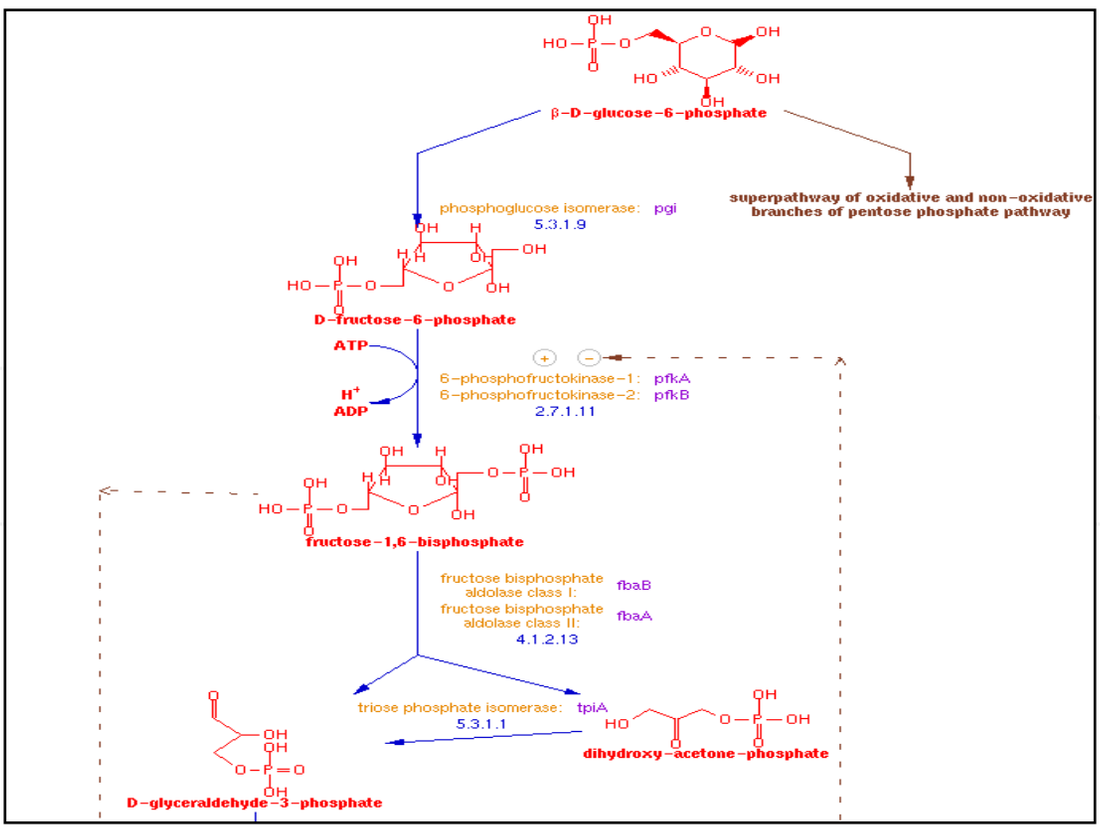
Data integration aims to work with repositories of data from a variety of sources. As such, two databases may not provide identical information, and integrating these two databases may yield a richer resource for analysis. The conditions under which data is collected and the supporting references play a crucial role in making the analysis more meaningful. So far, the integration approaches have focused on different types of pathways. The same pathway can have different representations in different databases.

For example, a known pathway like Glycolysis is represented in different ways in KEGG and BioCyc as shown in Figure 3. A universal tool to integrate all types of pathways may not be a focus. Additionally, different databases employ various data representations that may not provide easy user access or user friendly. Figure 3(a) and 3(b) illustrate representational difference between two data sources for the same pathway. Various data integration models are defined below.

- *Syntactic Networks*: Syntactic networks adhere to the syntax of a set of words as given by the representation of the data and do not interpret the meaning associated. Syntactic heterogeneity is a result of differences in representation format of data.
- *Semantic Networks* (SN): Semantic heterogeneity is a result of differences in interpretation of the 'meaning' of data. Semantic models aim to achieve semantic interoperability, a dynamic computational capability to integrate and communicate both the explicit and implicit meanings of digital content without human intervention.
- Several features of SN make it particularly useful for integrating biological data include, ability to easily define an inheritance hierarchy between concepts in a network format, allow economic information storage and deductive reasoning, represent assertions and cause effect through abstract relationships, cluster related information for fast retrieval, and adapt to new information by dynamic modification of network structures [44]. An important feature of SN is the ease and speed to retrieve information concerning a particular concept. The use of semantic relationships ensures clustering together related concepts in a network. For example, protein synonyms, functional descriptions, coding sequences, interactions, experimental data or even relevant research articles can all be represented by semantic agents, each of which is directly linked to the corresponding protein agent.



(a)



(b)

Figure 3. (a) Pathway from KEGG- Glycolysis (b) BioCyc- Glycolysis

Biological information can be retrieved effectively through simple relationship traversal starting from a query agent in the semantic network. Two approaches primarily in practice for SNs are;

1. memory-mapped data structure and
2. indexing flat files.

In the memory-mapped data structure approach, subsets of data from various sources are collected, normalized, and integrated in memory for quick access. While this approach performs actual data integration and addresses the problem of poor performance in the federated approach, it requires additional calls to traditional relational databases to integrate descriptive data. While data cleaning is being performed on some of the data sources, it is not being done across all sources or in the same place. This makes it difficult to quickly add new data sources. In the indexing flat files approach, flat text files are indexed and linked thus supporting fast query performance.

- *Causal Models*: A causal model is an abstract model that uses cause and effect logic to describe the behaviour of a system. Ex: Expression Quantitative Trait Loci: (eQTLs) eQTL analysis is to study the relationship between genome and transcriptome. Gene expression QTLs that contain the gene encoding the mRNA are distinguished from other transacting eQTLs. eQTL mapping tries to find genomic variation to explain expression traits. One difference between eQTL mapping and traditional QTL mapping is that, traditional mapping study focuses on one or a few traits, while in most of eQTL studies, thousands of expression traits get analyzed and thousands of QTLs are declared.
- *Context likelihood of relatedness* (*CLR*): It uses transcriptional profiles of an organism across a diverse set of conditions to systematically determine transcriptional regulatory interactions. *CLR* is an extension of the relevance network approach. (http://gardnerlab.bu.edu/software&tools.html). [34] Presented architecture for context-based information integration to solve semantic difference problem, defined some novel modeling primitives of translation ontology and propose an algorithm for translation.
- *Bayes Networks (BN)*: Probabilistic graphical models that represent a set of variables and their probabilistic independencies. For example, a BN could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Bayes networks focus on score-based structure inference. Available heuristic search strategies include simulated annealing and greedy hill-climbing, paired with evaluation of a single random local move or all local moves at each step. [45] Bases his approach on the well-studied statistical tool of Bayesian networks [46]. These networks represent the dependence structure between multiple interacting quantities (e.g., expression levels of different genes). His approach, probabilistic in nature, is capable of handling noise and estimating the confidence in the different features of the network.
- *Hidden Markov Models (HMM)*: HMM is a statistical model that assumes the system being modeled to be a Markov process with unknown parameters, and determines the

hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications. An HMM can be considered as the simplest dynamic Bayesian network. HMMs are being applied to the analysis of biological sequences, in particular DNA since 1998 [47].

3.2. Need to use open grid service architecture ogsa-dai for data access and integration

Apart from the ubiquitous call for more functionality, bioinformatics projects with commercial users/partners are very anxious about the security of their data. The issue is further complicated by the lack of coherent security models with the evolving WS-RF and WS-I specifications which OGSA-DAI now supports. This issue needs to be resolved if bioinformatics projects with commercial users/partners are not to be deterred from adopting the product despite its utility. In contrast to the diversity of its data resources, a limited range of operations on these resources is typically required. For instance, one operation is to create a study data set by aggregating data from iterative searches of remote data collections using the same taxonomy object (representing a species or other group) as the search parameter [48].

3.3. Handling the heterogeneity in data representation among databases

For biological plant pathways, various databases incorporate information about an entity/reaction/pathway to a level of detail and define their own data format. This includes information like number of fields, column label/tag, pathway name(s), etc. At the outset, common information across the tables may look limited and hard to extract mainly because of the tag or synonyms (other names) of pathway. Before proceeding for integration of a pathway across data sources following steps need to be carried out. For biological pathway integration, following needs to be considered.

- What is the aim of integration?
  To query autonomous and heterogeneous data sources through a common, uniform schema (TARGET SCHEMA).
- How will the integrated data be used?
  - Resolving various conflicts between source and target schema.
  - Offering a common interface to access integrated information.
  - Preserving the autonomy of participating systems.
  - Easily integrating data sources without major modification.
- Is it within a single data source or across sources?
- Does it support web based integration?
- Does it encompass the dynamic nature of the data?
- What are the data, source, user models, and assumptions underlying the design of integration system?

Specific data integration problems in the biological field include:

- Some biological data sources do not provide an expressive language
- Derived wrapper (operate in two modes)
  - Traditional wrapper
  - Virtual source that buffers the execution result of a local application
- Data Model Inconsistencies requires complex data transformation coding
- Data Schema Inconsistencies
- Schema matching: error-prone task
- Mapping info: systematically managed
- Domain Expert participation
- Along with the data schema consistencies there may be data level inconsistencies such as:
- Data conflict as each object has its own data type, and may be represented in different formats
- Different Query Capabilities affect the query optimization of data integration system
- Miscellaneous: Network environment, Security

File formats: For biological pathways, various data sources incorporate information about an entity/reaction/pathway to a level of detail and define their own data format. This includes information like number of fields, column label/tag, pathway name(s), etc. At the outset, common information across the tables may look limited and hard to extract mainly because of the tag or synonyms (other names) of pathway. One of the other important differences in the way these data sources are developed lies in the synonym representations. Some of the data sources limit the synonyms to 10 others may not result into may be over 40 synonyms. While we look at the data integration mechanism, if the names of the compounds do not match, then the search should be carried forward with the list of synonyms. In integrating different data bases this will take different search time. Also, since the field names (compound names) did not match, the search must unify the field names and generate a new list of synonyms.

Granularity of information: Different pathway databases may model pathway data with different levels of details. This primarily depends on the process definition. For example, one database might treat processes together as a single process, while another database might treat these as separate processes. Also, one database might include specific steps to be part of the process, while another database might not consider these steps. Additionally, the levels of details associated with a certain data base necessitate pathway data modeling with different levels of granularity. Different pathway data formats (e.g., SBML and BIND XML) have been used to represent data with different levels of details. A semantic net based approach to data integration is proposed in [49].

Heterogeneous formats: As the eXtensible Markup Language (XML) has become the lingua franca for representing different types of biological data, there has been a proliferation of semantically-overlapping XML formats that are used to represent diverse types of pathway data. Examples include the XML-derivatives KGML, SBML, CellML, PSI MI, BIND XML, and Genome Object Net XML. Efforts have been underway to translate between these

formats (e.g., between PSI MI and BIND XML, and between Genome Object Net and SBML). However, the complexity of such a pair-wise translation approach increases dramatically with a growing number of different pathway data formats. To address this issue, a standard pathway data exchange format is needed. While the Resource Description framework (RDF) is an important first step towards the unification of XML formats in describing metadata (ontologies), it is not expressive enough to support formal knowledge representation [50]. To address this problem, more sophisticated XML-based ontological languages such as the Web Ontology Language (OWL) have been developed. An OWL-based pathway exchange standard, called BioPAX, has been released to the research community [51].

4. Biological Pathway Data Integration

An integration model may serve as a tool to the user for a specific type of pathway. An algorithm for integration is presented next.

Metabolic Pathways: Integrating pathways from different data sources for the same species extract similar structures in them as the first step; this step integrates vertically given pathway within a species across data sources. (Database is the variable) this includes sorting a graph $G\ (V, E)$ for common V's and E's in $G_i\ (V_i, E_i)$ and $G_j\ (V_j, E_j)$. In the discussion that follows, *i*ntegrating pathway as the TCA cycle given by two data sources namely; *KEGG* $(D^{ij}{}_1)$ and BioCyc $(D^{ij}{}_2)$ for *E. coli K-12* is considered. For metabolic pathways the details associated with each graph include the nodes and edges as given below. For Protein-Protein interaction the nomenclature and associated fields for nodes and edges may change. However, it is possible to come up with a structure that can describe the Protein-Protein interactions or signal transduction pathways.

- *Node:* Biological Name, ID, Neighbor, Type, Context, Pathway, Data Source, PubList, SynList, empirical formula, Structure
- *Edge:* EdgeID, EdgeSource, EdgeDest, Reactiontype (Rev/ Irreversible), Data Source, Enzyme, Genes

Signal Transduction Pathways: The information contained in signal transduction pathways is not similar to the metabolic pathways. In signal transduction pathways, the interactions can be represented as a class hierarchy. Our aim is also to integrate a sample pathway like insulin from sources like KEGG, SPAD to see the performance of our algorithm. Interestingly, SPAD assigns evidence code to the edges (interactions) and KEGG assigns only one evidence code to the pathway (nodes and edges). The format of the table for integration is given above. Before integration information associated with every object (node) and edge (interactions) should be considered.

Before proceeding for integration of a pathway across data sources following steps need to be carried out.

Step 1

- Check for the pathway name across the input pathways.

- If a synonym matches then, go to step 2 else, search for synonyms of pathway name.

Step 2

- Choose the integrated output table format as the reference (number of columns, column tag)
- Check for number of columns in the output table.
- Match each of the column names in the output table with each of the column names in the input data files,
  - if column names are same then continue, else see alternate tag for the column, and match them.
- Match order in the output table format with the inputs from different sources.
  - If the order matches, then continue, else reorder the columns as given in output table.
- Check for number of columns in the output table,
  - If the number of columns is not same, then append the table with new columns.

Step 3

- Apply EV and Integration algorithms

The notations used in our algorithm are presented next.

4.1. Notations

- $\boldsymbol{S} = \{s^1, s^2, s^3, \ldots s^n\}$ is set of species. (1)
- $\boldsymbol{P^{ij}} = \{p^{i1}, p^{i2}, \ldots p^{ip}\}$ is a set of pathways within s^i (2)

Consider a tuple $(S^i, (P^{ij}, (D^{ij}{}_i))$ (3)

Where, $\boldsymbol{D^{ij}{}_k} = \{d^{ij}{}_1, d^{ij}{}_2, d^{ij}{}_3, \ldots d^{ij}{}_k\}$ is a set of 'k' data sources for (S^i, P^{ij}) (4)

- $s^1 = \{(s^1, p^{1j} (D^{1j}{}_k)\} = \{(s^1, p^{1j}, d^{1j}{}_1\}) (s^1, p^{1j}, d^{1j}{}_2\}), \ldots (s^1, p^{1j}, d^{1j}{}_k)\}$ for 'k' databases,

For example; s^1: *E.coli*; p^{1j}: TCA Cycle; $d^{1j}{}_{1=}$BioCyc, $d^{1j}{}_{2=}$KEGG.
Then, the tuple $(v^{11}{}_{1n}, e^{11}{}_{1m})$ gives *(node, edge)* in Biocyc for TCA cycle in *E.coli*, and the tuple $(v^{11}{}_{2p}, e^{11}{}_{2p})$ gives *(node, edge)* in KEGG for TCA cycle in *E.coli*

- $s^2 = \{(s^2, p^{2j} (D^{2j}{}_k)\} = \{(s^2, p^{2j}, d^{2j}{}_1), (s^2, p^{2j}, d^{2j}{}_2), \ldots\ldots, (s^2, p^{2j}, d^{2j}{}_r)\}$ *for 'r'* databases,

For example, s^2: *Arabidopsis*; p^{2j}: TCA Cycle; $d^{2j}{}_{1=}$BioCyc, $d^{2j}{}_{2=}$ AraCyc
Then, the tuple $(v^{22}{}_{1p}, e^{22}{}_{1p})$ gives the *(node, edge)* in AraCyc for TCA cycle in *Arabidopsis*, and the tuple $(v^{22}{}_{2p}, e^{22}{}_{2p})$ gives the *(node, edge)* in KEGG for TCA cycle in *Arabidopsis.*

Each pathway p^{ij} *for a* $d^{ij}{}_k$ is given by a graph $G\ (V^{ij}{}_k, E^{ij}{}_k)$, where,

- $\boldsymbol{P^{ij}{}_k} = \boldsymbol{G\ (V^{ij}{}_k, E^{ij}{}_k)}$ represents Pathway *'j'* from k^{th} datasourcesS for species *i'*… .(5)

Where, $\boldsymbol{V^{ij}{}_k} = \{v^{ij}{}_{k1}, v^{ij}{}_{k2}, \ldots v^{ij}{}_{kn}\}$ = set of nodes in $d^{ij}{}_k$,. (6)

$\boldsymbol{E^{ij}{}_k} = \{e^{ij}{}_{k1}, e^{ij}{}_{k2}, \ldots . e^{ij}{}_{km}\}$= set of edges in $d^{ij}{}_k$,. (7)

- ***SynList*** *{pathway name}* = ***SynList*** $\{P^{ij}\}$

- ***SynList*** *{entity name}* **=** ***SynList*** *{v^{1j}_{kn}}*
- ***EV***$^{ij}_{k}$ = *{EV^{ij}_{k1}, …. EV^{ij}_{kh}}* set of *'h'* EV Codes for *{s^i, p^{ij} , d^{ij}_k }*, for example;
  - EV^{1j}_1= *{*Set of EVcodes given by Biocyc for E.coli for TCA cycle*}*
  - EV^{1j}_2= *{*Set of EV codes given by KEGG for E.coli for TCA cycle*}*
  - EV^{2j}_3= *{*Set of EV codes given by AraCyc for Arabidopsis for TCA cycle*}*
  - EV^{2j}_2= *{*Set of EV codes given by KEGG for Arabidopsis for TCA cycle*}*
- ***RI***$^{ij}_{k}$*:* Reference index for a database d^{ij}_k
- ***RI***$^{ij}_{int}$*:* Reference index for the integrated pathway
- ***CW***$^{ij}_{k}$*:* Confidence weight for a database d^{ij}_k
- ***CW***$^{ij}_{int}$*:* Confidence weight of the integrated pathway p^{ij} within a species
- ***V***$^{ij}_{int}$*:* Integrated node table for a species S^i, for a pathway p^{ij}
- ***E***$^{ij}_{int}$*:* Integrated edge table for a species S^i, for a pathway p^{ij}
- *(v^{1j}_{kn}, e^{ij}_{km}) = (node 'n', edge 'm')* in d^{1j}_k of s^1 for p^{1j};
- ***ATT*** *{(v^{1j}_{kn} ,(A)}= {v^{1j}_{kn}, (A_1, A_2, A_3, A_4, …A_s)}* = set of attributes of the node v^{1j}_{kn}
- ***ATT*** *{(e^{ij}_{km}, (B)} ={(e^{ij}_{km}, (B_1, B_2, B_3, … B_t)}* = set of attributes of edge e^{ij}_{km}
- ***DATT*** *{v^{1j}_{kn} ,(δA)}* = set of derived attributes of the node v^{1j}_{kn} *(EV_i, CW_i, RI_i)*
- ***DATT*** *{e^{1j}_{kn} ,(δB)}*= set of derived attributes of the edge e^{1j}_{kn} *(EV_i, CW_i, RI_i)*
- ***δV***$^{ij}_{k}$ = Set of derived node attributes for Integrated pathway *{EV_{int}, CW_{int}, RI_{int}}*
- ***δE***$^{ij}_{k}$= Set of derived edge attributes for Integrated pathway *{EV_{int}, CW_{int}, RI_{int}}*
- ***V***$^{ij}_{int}$ = *{*Σ ***V***$^{ij}_{k}$ *}* *for k= 1 to n*
- ***E***$^{ij}_{int}$= *{*Σ ***E***$^{ij}_{k}$*}* *for k= 1 to n*
- ***P***$^{ij}_{int}$ = Integrated pathway from multiple DSs = *{*Σ ***P***$^{ij}_{k}$ *}* *for k=1 to n*

4.2. Biological Pathway Data Integration Algorithm

Following selections and inputs are defined by the user.

- User selected inputs: Species, Pathway, Data sources/database
- User inputs: Confidence assigned to each database
- User defined filters (**UDF**) for entities like substrate nodes, H_2O, CO_2 etc. for integrated pathway [$P^{ij}_{int} = G\ (V_{int}, E_{int})$],

Step 1.

For each user selected pathway ***P***^{ij} *for a species* s^i

List D^{ij} (d^{1j}_1,… d^{nj}_k), ***(KEGG, BioCyc, MetNetDB etc)***

Step 2. **Define** rules to classify the interactions, for example;

- If the pathway is *signal transduction*, then use the *classifier (Table 1)*for interactions
- If the pathway is *metabolic,* then *reaction* is a general representation of the interaction

Sort (d^{1j}_1,… d^{nj}_k) according to species (s^i,d^{ij}_1), (s^j,d^{ji}_1) etc.

Generate a set of (nodes, edges) from all the input data sources *{(V^{ij}, E^{ij})} = {(V^{ij}_1, E^{ij}_1), (V^{ij}_2, E^{ij}_2)….. (V^{ij}_s, E^{ij}_s)}*

where, $V^{ij}_k = \{v^{ij}_{k1}, v^{ij}_{k2}, \ldots v^{ij}_{kt}\}$ and $E^{ij}_1 = \{e^{ij}_{k1}, e^{ij}_{k2}, \ldots e^{ij}_{ku}\}$

Step 3.

For $k = 1, \ldots, q$ $(d^{1j}_1, \ldots d^{1j}_k)$,
For $s = 1, .., n$, and $q = 1, \ldots m$,
List ***ATT*** $\{(v^{1j}_{ks}, (A)\}$
List ***ATT*** $\{(e^{ij}_{kq}, (B)\}$
Select $v^{ij}_{k1} \in V^{ij}_k$ **C** d^{1j}_k
For all $p = 1$ *to* n
Check *for* $v^{ij}_{k,1} \in V^{ij}_p$ (node name match across data sources)
If *YES,* then **Apply** *EV* integration algorithm
Generate ***DATT*** $\{v^{1j}_{kn}, (\delta A)\}$, ***DATT*** $\{e^{1j}_{kn}, (\delta B)\}$,
Else, *For* $p = 1$ *to* n,
For $t = 1, z$
Check if $v^{ij}_{k,1} \in$ ***SynList*** $\{v^{ij}_{p,t}\}$ (node name(A) with Synlist(B))
If YES, then Apply EV integration algorithm,
Generate ***DATT*** $\{v^{1j}_{kn}, (\delta A)\}$, ***DATT*** $\{e^{1j}_{kn}, (\delta B)\}$,
Else,
Check if ***SynList*** $\{v^{ij}_{k,1}\}$ has a match with $v^{ij}_{p,t}$
If YES, then Apply EV integration Algorithm
Else,
Check if ***SynList*** $\{v^{ij}_{k,1}\}$ has a match with ***SynList*** $\{v^{ij}_{p,t}\}$
If $v^{ij}_{k,1} = v^{ij}_{p,l}$ *is TRUE,*
Then,
Include $v^{ij}_{k,1}$ with the matched node name $v^{ij}_{k-1,p} \in V^{ij}_{k-1}$
Compute $(\delta V^{ij}_k, \delta E^{ij}_k)$

This is the node name for the integrated database for the species. **Level 1**

Generate SynListInt = {**SynList** $(v^{ij}_{k,1})$ U ***Synlist*** $(v^{ij}_{p,l})$U…} without duplication
Associate DOI (date of integration)
Generate P^{ij}_{int}

$$\boldsymbol{P^{ij}_{int}} = \{\Sigma\, \boldsymbol{P^{ij}_k}\}_{\text{for } k=1 \text{ to } n} = [\{\boldsymbol{V^{ij}_{int}}, \boldsymbol{E^{ij}_{int}}\} + \{\Sigma\, \boldsymbol{\delta V^{ij}_{kt}}, \Sigma\, \boldsymbol{\delta E^{ij}_k}\}_{\text{for } k=1 \text{ to } n}]_{at\ t=t1}$$
$$= \Sigma\, \{\boldsymbol{ATT}\,[(v^{1j}_{kn}, (A)]\}, \boldsymbol{ATT}\,[(e^{ij}_{km}, (B)]\} + \Sigma\, \{\boldsymbol{DATT}\,\{v^{1j}_{kn}, (\delta A), \boldsymbol{DATT}\,\{e^{1j}_{kn}, (\delta B)\}_{for\ all\ n,\ m}\ \{\boldsymbol{\delta V^{ij}_k}\ \boldsymbol{\delta E^{ij}_k}\}$$

Step 4.

Repeat *Step 2-3 for* $e^{ij}_k \in E^{ij}_k$ *in* $(d^{ij}_1, \ldots d^{ij}_k)$, *for* p^{ij}
Include information associated with the edge, as given by 'edges' such as reaction, enzyme, by products and substrates along with attributes like evidence, reference publications, context etc.
** Outputs E^{ij}_{int} table for s^i using $(d^{1j}_1, \ldots d^{1j}_k)$, with EV^{ij}_{int}, CW^{ij}_{int} and RI^{ij}_{int}. **Level 1.** **

Step 5.

Generate integrated pathway by consolidating outputs $G(V^{ij}_{int}, E^{ij}_{int})$ for s^i

Step 6.

For i =1,…n

Repeat steps 2- 4 to integrate P^{ij} for all species s^i

** **This generates Table** $(V^j_{int}, E^j_{int}) = \{(V^{ij}_{int}, E^{ij}_{int}) \cup (V^{jj}_{int}, E^{jj}_{int}) \cup \ldots\}$ **for** S^i**, for all** $i = 1, \ldots n$**, for a** p^{ij}**. Level 2*****

Step 7.

For (j = 1,….p)

Integrate for all P^{ij}

This **generates output table $(V_{int}, E_{int}) = (V^j_{int}, E^j_{int}) \cup (V^k_{int}, E^k_{int}) \cup \ldots$**for all** $(j = 1,\ldots p)$**. Level 3*****

Step 8.

Apply UDF (User defined filter)

5. Querying Integrated Pathway

Once the data integration is accomplished, extracting information from the integrated data will be of interest to the biologist. There are various mechanisms to extract information from the integrated database generated. Some of these are described below.

Granular computing with semantic network structure captures the abstraction and incompleteness associated with biological plant pathway data. It is inspired by the ways in which humans granulate information and reason with coarse grained information. The three basic concepts underlying the human cognition are granulation, organization, and causation. Granulation involves decomposition of whole into parts, organization involves integration of parts into whole, and causation involves associations of cause and effects. The fundamental issues with granular computing are granulation of the universe, description of granules, and relationships between granules. The basic ideas of crisp information granulation have appeared in related fields, such as interval analysis, quantization, rough set theory, Demster Shafer theory of belief functions, divide and conquer, cluster analysis, machine learning, data bases and many others. Granules may be induced as a result of 1) equivalence of attribute values, 2) similarity of attribute values, and define the granules 3) equality of attribute value. We use granules for defining the user queries associated with the integrated pathway. Based on user (biologist) choice, granules can be defined to view the integrated pathway. This provides flexibility to the biologist for using the information.

Previous approaches towards metabolic network reconstruction have used various algorithmic methods such as name-matching in IdentiCS [52] and using EC-codes in metaSHARK [53] to link metabolic information to genes. The AUtomatic Transfer by Orthology of Gene Reaction Associations for Pathway Heuristics (AUTOGRAPH) method

[54] uses manually curated metabolic networks, orthologue and their related reactions to compare predicted gene-reaction associations.

Arrendondo [55] Proposes to develop a process for the continuous improvement of the inference system used, which is applicable to any such data mining application. It involves the comparison of several classifiers like Support Vector Machines (SVMs), Human Expert generated Fuzzy, and Genetic Algorithm (GA) generated Fuzzy and Neural Networks using various different training data models. In his approach, all classifiers were trained and tested with four different data sets: three biological and a synthetically generated mixture data set. The obtained results showed a highly accurate prediction capability with the mixture data set providing some of the best and most reliable results.

6. Conclusion

Biological database integration is a challenging task as the databases are created all over the world and updated frequently. For biological data sources that may be derived from an earlier existing data source, it is also important to identify the evidence of the data source represented by the evidence code, to be included as a candidate for integration. In most data integration algorithms the user does not participate thus leading to an integrated data source with any effective utility towards analysis.

Large scale integration of pathway databases promises to help biologists gain insight into the deep biological context of a pathway. In this chapter, we presented algorithms that help user to select their choice of data sources and apply Evidence code algorithm to compute an integrated EV code and RI for the pathway data of interest. The ultimate goal is to generate a large-scale composite database containing the entire metabolic network for an organism. This qualitative approach includes aspects like user confidence scores for databases for mapping EV and generating RI for a given pathway. For the TCA pathway results show that generating such a mapping is helpful in visualizing the integrated database that highlights the common entities as well as the specifics of each database. As the database confidence weight selection is user specific, the integration yields different results for different users for the same database which will allow users to explore the effects of different hypotheses on the overall network. Once the integrated evidence code is generated, then data integration algorithm is applied to get the integrated pathway data. To best attempt integration of such data it is imperative to include user participation as user mostly identifies the associations and behavior of various compounds, reactions, genes in a given biological pathway leading to significant diagnosis.

Author details

Shubhalaxmi Kher
Electrical Engineering, Arkansas State University, USA

Jianling Peng
Samuel Roberts Noble Foundation, USA

Eve Syrkin Wurtele
Department of Genetics, Development and Cell Biology, Iowa State University, USA

Julie Dickerson
Electrical and Computer Engineering, Iowa State University, USA

7. References

[1] Akula, S.; Miriyala, R.; Thota, H.; Rao, A.; Gedela, S. Techniques for Integrating –omics Data, Bioinformation, Views and Challenges, 2009.
[2] Saccharomyces genome database. http://www.yeastgenome.org/
[3] KEGG: Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jp/kegg/
[4] TAIR- AraCyc: http://www.arabidopsis.org/biocyc/
[5] Thimm, O; Blasing, O; Gibon, Y; Nagel, A; Meyer, S; Kruger, P; Selbig, J; Muller, L; Rhee, S; and Stitt, M. MAPMAN: a user driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes, The Plant journal (2004) 37, pp 914-939. http://www.uky.edu/~aghunt00/PLS620/papers.htm/Systems%20approaches%20copy/MAPMAN.pdf
[6] BIND: Biomolecular Interaction Network Database http://metadatabase.org/wiki/BIND_-_Biomolecular_Interaction_Network_Database
[7] Bajic VB, Veronika M, Veladandi PS, Meka A, Heng MW, Rajaraman K, Pan H, Swarup S. Dragon Plant Biology Explorer. A text-mining tool for integrating associations between genetic and biochemical entities with genome annotation and biochemical terms list, Plant Physiol. 2005 Aug; 138(4):1914-25.
[8] Pandey R, Guru R K, Mount D W. Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data, Bioinformatics. 2004 Sep 1;20(13):2156-8. Epub 2004 May 14.
[9] RegulonDB database: Escheichia Coli k-12 transcriptional network. http://regulondb.ccg.unam.mx/
[10] PlantCare a database. http://bioinformatics.psb.ugent.be/webtools/plantcare/html/
[11] PLACE: a database of Plant Cis-acting regulatory netowrk. http://www.dna.affrc.go.jp/PLACE/
[12] EPD: Eukaryotic promoter database. http://epd.vital-it.ch/
[13] TRRD: transcription regulatory regions database http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/
[14] Athamap: http://www.athamap.de/
[15] TRANSFAC: http://www.gene-regulation.com/pub/databases.html/
[16] Friedman N, Linial, M; Nachman,I; and Pe'er, D. Using Bayesian Networks to Analyze Expression Data, Journal of computational biology, Volume 7, Numbers 3/4, 2000, pp. 601-620.
[17] Schadt, et.al, *An Integrative Genomics Approach to Infer Causal Associations Between Gene Expression and Disease*, *Nature Genetics, vol.37, number 7, July 2005, pp, 710-717*.
[18] The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/).

[19] Liu, Y.; Wang, Y.; Liu, Y.;, Tan, Z. Data Integration of Bioinformatics Database Based on Web Services, International Journal of Web Applications, Volume 1, Number 3, 2009.
[20] UCLA-DOE Institute for Genomics and Proteomics. http://dip.doe-mbi.ucla.edu/dip/Main.cgi
[21] IntAct: open source database system and analysis tools for molecular interaction data.http://www.ebi.ac.uk/intact/
[22] GRID: http://www.moldiscovery.com/soft_grid.php/
[23] Zanzoni, A. Montecchi-Palazzi, L. Quondam, M. Ausiello, G. Helmer-Citterich, M. Cesareni, G. *MINT: A Molecular INTeraction database. Elsevier FEBS Letters, 2002, Volume 513, Issue 1, Pages 135-140.*
[24] Coessens B. et.al, *INCLUSive: A Web Portal and Service Registry for Microarray and Regulatory Sequence Analysis, Nucleic AciDS research, 2003, vol. 31, No.13. pp. 3468-3470.* http://tomcatbackup.esat.kuleuven.be/inclusive/
[25] Achard, F.; Vaysseix, G.; Barillot, E. XML, Bioinformatics, and Data Integration, Bioinformatics Review, Evry, France, 2001, pp. 115-125.
[26] Pathway Data List. http://cbio.mskcc.org/prl
[27] Hsing, M., Cherkasov, A. Integration of Biological Data with Semantic Networks, Current Bioinformatics, 2006, 1 000-000.
[28] Chung, M., Lim, M., Bae, M., Park, S. Customized Biological Database Integration for cDNA Microarray, RECOMB 2005, Research in Computational and Molecular Biology, Cambridge, 2005.
[29] Gopalcharyulu, P. Lindfors, E. et.al. Data integration and visualization system for enabling conceptual biology, BioInformatics, Vol.21, Suppl 1 2005, pp. i177-i185.
[30] Rzhetsky, A et.al, GeneWays: A System for Extracting, Analyzing, Visualizing and Integrating Molecular Pathway Data, Journal of Bioinformatics, 2004, 43-53.
[31] Zucker, J.,Luciano, J., Brandes, A. Lin, X. Semantic Aggregation Integration and Inference: Three case studies, ISMB 2005.
[32] Hu, Z., Mellor, J., Wu, J., Yamada, T., Holloway, D., DeLisi, C. VisANT: Data integrating visual framework for biological networks and modules, Nucleic AciDS research, 2005 vol. 33.
[33] Zhang Z.; Bajic, V.; Yu, J.; Cheung, K.; Townsend, J. Data Integration in Bioinformatics: Current Efforts and Challenges. Bioinformatics: Trends and Methodologies, Intech, 2011.
[34] Zhang, D. and Jing, L., *Context based Numerical information, IEEE conference on E-commerce Technology 2005*Arredondo, T., Seeger, M., Dombrovskaia, L., Avarias, J., Calderón, F., Candel, D., Muñoz, F., Latorre, V., Agulló, L., Cordova, M., and Gómez, L.: "Bioinformatics Integration Framework for Metabolic Pathway Data-Mining". In: Ali, M., Dapoigny, R. (eds): Innovations in Applied Artificial Intelligence. Lecture Notes in Artificial Intelligence, Vol. 4031. Springer-Verlag, Berlin (2006) pp. 917-926.
[35] PATIKA: http://www.iam.metu.edu.tr/research/groups/compbio/PATIKA_METU04.pdf
[36] INHO: http://www.inoh.org/
[37] TRANSPATH. http://www.ncbi.nlm.nih.gov/pubmed/12519957
[38] ReactomeSTKE. http://stke.sciencemag.org/

[39] MetaCyc. http://metacyc.org/
[40] Kher, S; Jianling Peng; SyrkinWurtele, E.; Dickerson, J. A Symbolic computing approach to evidence code mapping for biological data integration and subjective analysis for reference associations for metabolic pathways, Annual Meeting of the North American Fuzzy Information Processing Society, 2008, NAFIPS 2008. NY 2008. pp. 1-6.
[41] Kher, S; Dickerson, J; Rawat N. Biological pathway data integration trends, techniques, issues and challenges: A survey, Nature and biologically inspired computing, NaBIC 2010, Second World Congress, Fukuoka, Japan, 2010, pp.177 – 182.
[42] MetNetDB. http://www.metnetdb.org/MetNet_db.htm
[43] Karp, P. D., Paley, S., Krieger, C. J. An Evidence Ontology for Use in Pathway/Genome DS, Pacific Symposium on Biocomputing 2004, pp. 190-201, Singapore Bounsaythip, C., Lindfors, E., Gopalacharyulu, P., Hollmen, J., and Oresic, M. *Network Based Representation of Biological Data for Enabling Context Based Mining, Bioinformatics, vol.21, suppl 1. 2005, pp. 177-185.*
[44] Newman, M. E. J and Leicht, E. A *Mixture Models and Exploratory Analysis in Networks, Physics, May 2007*.
[45] Pearl. J. (2000) *Causality: Models, Reasoning, and Inference.*Cambridge University Press, 2000.
[46] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, USA: Morgan Kaufmann Publishers.
[47] Christopher Nemeth, STOR-I, Hidden Markov Models with Applications to DNA Sequence Analysis.
[48] Crompton, S.; Matthews, B.; Gray, A.; Jones, A.; White, R. Data Integration in Bioinformatics Using OGSA-DAI, In Proceedings of Fourth All Hands Meeting, 2005.
[49] Cheung Kei-hoi; Qi, P; Tuck,D; Krauthammer,M. A Semantic Web Approach to Biological Pathway Data Reasoning and Integration, Elsevier Vol. 4, issue 3, Sep 2006, pp. 207-215.
[50] RDF-OWL. http://www.w3.org/RDF/
[51] BioPAX: http://www.biopax.org/
[52] Sun, J. and Zeng, A. , IdentiCS – Identification of coding sequence and *in silico* reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence, *BMC Bioinformatics* 2004, 5:112 doi:10.1186/1471-2105-5-112
[53] Pinney, J.W., Shirley, M.W., McConkey, G.A., Westhead, D.R. (2005) MetaSHARK: software for automated metabolic network prediction from DNA sequence and is application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*, *Nucleic Acids Research*, 33, 1399-1409.
[54] Notebaart, R. A., F. H. van Enckevort, C. Francke, R. J. Siezen, and B. Teusink. 2006. Accelerating the reconstruction of genome-scale metabolic networks. BMC Bioinformatics 7:296
[55] Arredondo, T., Seeger, M., Dombrovskaia, L., Avarias, J., Calderón, F., Candel, D., Muñoz, F., Latorre, V., Agulló, L., Cordova, M., and Gómez, L.: "Bioinformatics Integration Framework for Metabolic Pathway Data-Mining". In: Ali, M., Dapoigny, R.(eds): Innovations in Applied Artificial Intelligence. Lecture Notes in Artificial Intelligence, Vol. 4031. Springer-Verlag, Berlin (2006) p. 917-926