Integration of large datasets for plant model organisms

by

Yves Sucaet

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Eve Syrkin Wurtele, Co-Major Professor
Julie A. Dickerson, Co-Major Professor
Basil Nikolau
Shashi K. Gadia
David Fernandez-Baca

Iowa State University
Ames, Iowa
2013

DEDICATION

This work is dedicated to my daughter, Lily Claire Sucaet.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It is a pleasure to express my gratitude toward the many people who made this thesis possible. I'd like to thank my parents first, André Sucaet and Juliette Wynant, who taught me that I can be anything I want to be, if I'm willing to work hard enough for it. Nil volentibus arduum.

I am thankful to my supervisor, Dr. Eve Syrkin Wurtele, who supported me in a number of ways. Her encouragement, sound advice, guidance and lots of good ideas enabled me to develop and grow throughout my stay at Iowa State University. Her enthusiasm, inspiration, and great efforts to explain things clearly and simply, helped to make bioinformatics fun for me.

Thanks to my co-major professor Dr. Julie A. Dickerson. I would have been lost without her. I am indebted to my remaining committee members, Dr. Basil Nikolau, Dr. Sashi Gadia and Dr. David Fernandez-Bacha. Their kind assistance, wise advice, and help with various applications, has been invaluable. Special thanks go to Dr. Michael Stewart and Dr. Zhijun Wu, for influencing and helping to shape my thoughts early on in my graduate career.

I am very appreciative of my many student colleagues for providing a stimulating environment in which to learn and grow. I am especially grateful to Matthew Moscou, Misha Rajaram, Fadi Towfic, Michael Zimmermann, Preeti Bais, Xiaoyoung Sun, Sean Mooney, Xinyuan Zhao and many others. I would like to express my deepest gratitude to my wife, Tamera Elaine Sucaet, for taking a leap of faith and moving with me from

sunny Alabama to the cold winters of Iowa. All the members of the MetNet group and the Wurtele lab have been great companions along the road, especially Heather Babka and Nick Ransom.

I am grateful to the secretaries and librarians involved in the interdepartmental BCB program and GDCB department at Iowa State University, for helping the departments to run smoothly and for assisting me in many different ways. I'm particularly reminded of Trish Stauble, Constance Garnett, and Lynette McBirnie-Sprecher.

Lastly, I offer my regards to all of those who supported me in any respect during the completion of the project. Many people had minor, yet important, peripheral roles: Sven Aelterman, Bea Van Nieuwenhove, Taru Deva, Doug Hanson, Venkata Kishore, Rick Van Eeden, Onno Louwen, and Junli Yamada.

Support for this project was provided is in part by National Science Foundation Arabidopsis 2010 grant #052026. This work additionally supported by the National Science Foundation under Award No. EEC-0813570.

ABSTRACT

This dissertation is concerned with bioinformatics data integration. The first chapter illustrates the current state of biological pathway databases in general, and in particular, plant pathway databases. Key studies are cited to illustrate the potential benefits that may come from further research into integration methods.

Different models are explored to interface with the various stakeholders of biological data repositories. A public website (http://www.metnetonline.org) was built to address the role of a bioinformatics data warehouse as a server for external third parties. A dedicated API (MetNetAPI: http://www.metnetonline.org/api) accommodates bioinformaticians (and software developers in general) who wish to build advanced applications on top of MetNet. The API (implemented as .NET and Java libraries) was designed to be as user-friendly to programmers, as the public website is to end-users. Finally, a hybrid model is examined: the use of XML as a repository for information integration, downstream processing, and data manipulation. An overview of the use of XML in biological applications is included.

MetNetAPI functions according to certain principles; a subset of the API is abstracted and implemented to interface with a range of other public databases. This results in a new bioinformatics toolkit that can be used to mix and match data from heterogeneous sources in a transparent manner. An example would be the grafting of protein-protein interaction data on top of araCyc pathways.

Biological network data is often distributed over a variety of independently modeled databases. This dissertation makes two contributions to the field of bioinformatics: A new service – MetNet Online – is now operating which offers access to the earlier created and integrated MetNetDB data repository. The service is geared toward end-users, students and researchers alike, as well as seasoned bioinformatics software developers who wish to build their own applications on top of an already integrated datasource. Furthermore, integrated databases are only useful when they can be synchronized with their respective external sources. Thus, a framework was created that allows for a systematic approach to such integration efforts. In closing, this work provides a roadmap to maintain current as well as prepare for future integrated biological database projects.

CHAPTER I

PLANT PATHWAY RESOURCES AND DATABASES

Evolution and applications of plant pathway resources and databases

Plants are important sources of food and plant products are essential for modern human life. Plants are increasingly gaining importance as drug and fuel resources, bioremediation tools and as tools for recombinant technology. Considering these applications, database infrastructure for plant model systems deserves much more attention. Study of plant biological pathways, the interconnection between these pathways and plant systems biology on the whole has in general lagged behind human systems biology. In this article we review plant pathway databases and the resources that are currently available. We lay out trends and challenges in the ongoing efforts to integrate plant pathway databases and the applications of database integration. We also discuss how progress in non-plant communities can serve as an example for the improvement of the plant pathway database landscape and thereby allow quantitative modeling of plant biosystems. We propose Good Database Practice as a possible model for collaboration and to ease future integration efforts.

Introduction and background

A biological pathway is a programmed sequence of molecular events in a cell. This chain of events executes a particular cellular function or brings about a specific biological effect. Knowledge of an organism's pathways is essential to understand a

biological system at different levels, from simple metabolism to complex regulatory reactions. Many pathways are complex and hierarchical and are themselves interconnected to form, to participate in, or to regulate a network of events. Over the last couple of decades, there has been an exponential increase in the information on these pathways, their components and their functions [1]. This stems from the biotechnological advancements in genomics and proteomics and high throughput technologies like microarray and two-hybrid screens. For numerous species, this has increased our knowledge about normal pathways as well as rogue/aberrant pathways that lead to a variety of diseases. Examples include pathways that lead to cancer [2] or pathways that lead to aberrant leaf development in plants [3]. Production of large amounts of data necessitates the creation of pathway databases and repositories, where information about the pathways along with their molecular components and reactions is stored. These data sets often become data-sources in their own right, and are shared with the public, explaining in part the large number of databases that exist today [1].

Simultaneously, technological advancements that allow access to and discovery of novel pathway information have resulted in the creation of many more pathway databases [1] that target different organisms, processes and mechanisms. Availability of such vast amounts of information in an ordered format has led us to ask new questions. Ideker and colleagues [4] have raised questions pertinent to evolutionary and comparative biology, e.g. ‘considering that the protein sequences and structures are conserved, could the protein-interaction networks be conserved as well? Is there a minimal set of pathways that is required by all living organisms? Can the evolutionary

distance be measured at the network connectivity level rather than at the DNA or protein level?' Answers to these and other questions will lead to an increased understanding of living systems, which in turn may result in more questions, at other levels, that are currently unimaginable. Information aggregated from different pathway databases is often more useful than information from individual databases. Integration of information from various pathway databases can be used to reveal novel information about a system.

Information from pathway databases has been used for different purposes. Information analysis and data mining holds the potential for discovery of orthologous/analogous pathways and pathway components in other related organisms [5]. For example, organisms which are difficult to cultivate in vitro and therefore are less amenable to laboratory studies could be examined in silico through a study of orthologs. Iterative expansion of pathway data can be utilized to build models of biological mechanisms based on the hypotheses derived from these initial data; see Bumgarner and Yeung [6] for a recent review. Models can (and should) in turn generate experimentally verifiable predictions.

Pathway database analysis can be used to find patterns in the pathways that are related to a disease [7] and aid in the identification of new drug targets [8]. Another idea is targeted drug discovery by screening the complete pathway as compared to a single pathway component [9]. Pathway analysis can also be used to identify molecular switches that lead to disease and to efficiently turn them off to silence them without affecting the rest of the system. A recent study on riboswitches illustrates how one can reengineer components of a pathway to control expression of multiple genes [10].

Compared to the exponential increase in human/animal pathway databases, development of plant pathway databases has been modest and a smaller number of applications have resulted. Plant pathway databases have remained relatively under-utilized. This apparent lacuna is all the more concerning considering that plants are important as food crop, fiber and plant-based fuel source. Examples from non-plant resources and their applications can serve as inspiration for plant scientists who wish to control pathways, for instance, to produce crops with longer shelf life or enhance immunity to plant pathogens.

In this review, we provide an overview of existing plant pathway databases, look at current progress and how the information contained in the databases has been used in the past and can be used in the future. We use examples from the existing plant pathway databases to showcase the potential of database integration. Non-plant integration applications are discussed to suggest future potential. Finally, we discuss how already existing information can be further enriched, organized and utilized for practical applications. We also highlight the acute need of robust, long-term, and user-friendly interactive databases.

The pathway database landscape

Pathguide [1], an online pathway resource meta-database, provides an overview of more than 300 biological pathway resources that have been developed to date. These include pathway databases, tools for data analysis, visualization and data extrapolation and other (peripheral) databases that can be linked with pathway databases to provide

additional information. Some databases are specific to a particular organism, e.g. AraCyc [11] deals with the metabolic pathways of Arabidopsis thaliana. Some pathway databases are specific to a certain disorder or disease, e.g. the Human Cancer Protein Interaction Network (HCPIN)[12]; other contain information about a certain system in an organism, e.g. InnateDB [13], a repository for pathways involved in the innate immune system of humans and mice.

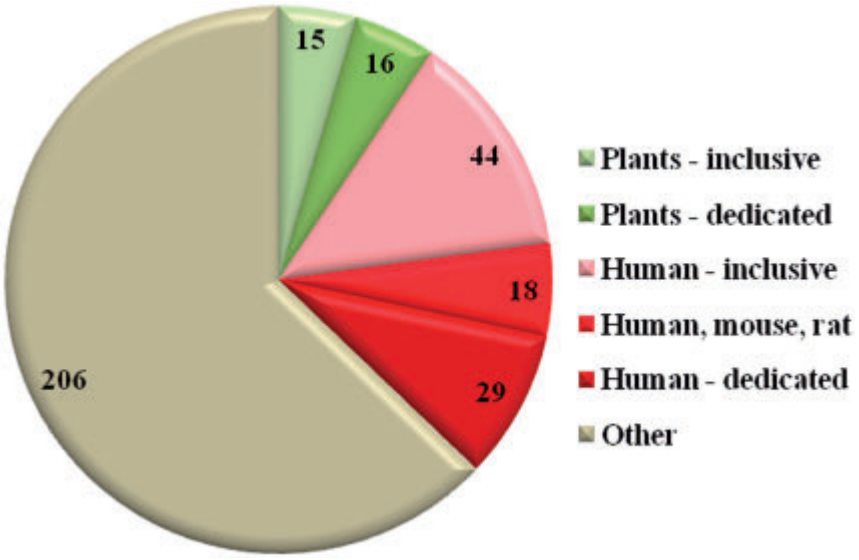

Figure 1 *- Pathway resources with plants and humans annotated asmajor organisms froma total of the 328 resources available in Pathguide. Inclusive - databases containing several other major organisms apart from plant or human; dedicated databases dedicated to plants or humans; other - databases for other organisms, or databases for numerous organisms which may also include human and plant information, and pathway tools. Numbers indicate the actual number of resources available for each category in Pathguide.*

Plant pathway databases, when compared to human pathway databases, are fewer in number (Figure 1) and much less diverse. There is an increasing awareness about the importance of plants as food crops, but it appears that only limited resources have been devoted to uncovering and understanding plant pathways. A comparison of the number of genomes sequenced to date for mammals and higher plants (Figure 2) shows that

plants receive less attention from the sequencing community when compared to other organisms. The absolute numbers differ between the databases (some sites are kept more current than others), but the trend remains the same. There are many biologically, medically and economically important plants that differ in their physiology. In addition, secondary metabolism is important from a pharmacological point of view. Therefore, there is a need for many more genomes to be sequenced, proteomes to be studied and pathways to be uncovered for the optimal utilization of plants. While lower numbers of genome sequencing data do not completely explain the lack of pathway databases, they certainly contribute to it.

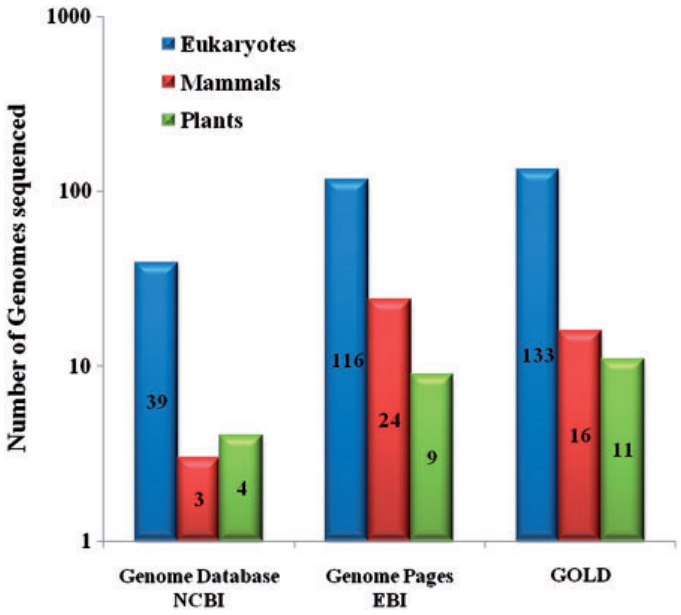

Figure 2: *A comparison of genomes sequenced for mammals and higher plants. Data from NCBI Genome Database (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html), Genome Pages at EBI (http://www.ebi.ac.uk/genomes/eukaryota.html) and GOLD database (http://www.genomesonline.org/cgi-bin/GOLD/bin/gold.cgi?page_requested¼CompletepPublished) are compared. Numbers in the bars indicate the number of genomes sequenced.*

Most plant pathway databases contain information on the networks in their own right, e.g. metabolic or regulatory networks in A. thaliana or soybean. However, there are no specialized databases yet that deal with pathways for plant immunity, plant growth or for controlling the size of plant organs.

For the purpose of this review, pathway databases are broadly classified into four types: metabolic pathways, gene regulatory networks, protein–protein interaction networks, and signaling pathways. 'Metabolic pathways' are the earliest discovered and best studied pathways.

Metabolic Pathways

Metabolic pathways are represented by a series of enzymatic reactions that take place at the level of small molecules. These have been elaborated and characterized for many organisms. Table 1 presents an overview of available metabolic pathway databases dedicated to different plant species and the sites that host them.

Organism	Database	Version	Location
Arabidopsis thaliana	AraCyc	7.0.0.0	http://www.arabidopsis.org/biocyc/index.jsp
	AraCyc	6.0.0.0	http://pathway.gramene.org/ARA/class-tree?object=Pathways
Oryza sativa japonica	RiceCyc	3.0.0.0	http://pathway.gramene.org/RICE/class-tree?object=Pathways
Sorghum bicolor	SorghumCyc	1.0.0.0	http://pathway.gramene.org/SORGHUM/class-tree?object=Pathways
Medicago truncatula	MedicCyc	1.0.1.1	http://pathway.gramene.org/MEDIC/class-tree?object=Pathways
Solanum lycopersicum	LycoCyc	2.0.1.1	http://pathway.gramene.org/LYCO/class-tree?object=Pathways
	LycoCyc	2.0.0.0	http://solcyc.solgenomics.net/LYCO/server.html
Capsicum	CapCyc	1.0.1.1	http://pathway.gramene.org/CAP/class-tree?object=Pathways
	CapCyc	2.1.0.0	http://solcyc.solgenomics.net/CAP/server.html
Solanum tuberosum	PotatoCyc	1.0.1.1	http://pathway.gramene.org/POTATO/class-tree?object=Pathways
	PotatoCyc	1.1.0.0	http://solcyc.solgenomics.net/POTATO/organism-summary?object=POTATO
Coffea canephora	CoffeaCyc	1.1.1.0	http://pathway.gramene.org/COFFEA/class-tree?object=Pathways
	CoffeaCyc	1.1.0.0	http://solcyc.solgenomics.net/COFFEA/organism-summary?object=COFFEA
Vitis vinifera	VitisNet		http://www.sdstate.edu/aes/vitis/pathways.cfm
Populus trichocarpa	PoplarCyc	2.0.0.0	http://pmn.plantcyc.org/POPLAR/server.html?
Petunia x hybrida	PetuniaCyc	2.1.1.0	http://solcyc.solgenomics.net/PET/server.html?
Solanum melongena	SolaCyc	1.2.0.0	http://solcyc.solgenomics.net/SOLA/organism-summary?object=SOLA
Nicotiana tabacum	NicotianaCyc	1.1.0.0	http://solcyc.solgenomics.net/TOBACCO/server.html?

Table 1: *Overview of plant species specific metabolic pathway databases*

Metabolic pathway databases like MetaCyc [14] contain experimentally verified metabolic pathways and enzyme information for more than 2000 organisms and can be used to predict orthologous pathways in another organism for which the genome has been sequenced and annotated. A dedicated portal for plant metabolic pathway databases is SolCyc (available at http://solcyc.solgenomics.net/). SolCyc is a Pathway Tools-based (and thus MetaCyc inferred) pathway genome database (PGDB) currently containing small molecule metabolism data for five plants belonging to family solanacea—tomato, potato, tobacco, pepper and petunia.

The pathways section of Gramene database [15] (a database for grasses such as rice, maize, sorghum, barley, oats, wheat and rye) contains the known and predicted biochemical pathways of rice (RiceCyc) and sorghum (SorghumCyc), both of which are curated by the Gramene database and were built using the Pathway Tools' PathoLogic module. The website also mirrors the known and predicted biochemical pathways from SolCyc, AraCyc, EcoCyc and the MetaCyc reference databases.

The 'golden standard' AraCyc for A. thaliana was built using the Pathway Tools' PathoLogic module with MetaCyc. AraCyc, in addition, uses manual curation to enrich its data. The trade-off is slower progress in completing the network, yet the end result is highly documented and has a more accurate structure. One can argue that databases are of higher quality when domain experts scrutinize the available literature and manually curate them. They can add their scientific experience and intuition to find facts in a way that any algorithm is yet to mimic. However, this all depends on the availability of such

experts and for genome-wide projects it is certainly challenging to gather all potentially involved.

The success of AraCyc has led to a broader plant-centric rather than organism-centric initiative, the Plant Metabolic Network (PMN) (available at http://www.plantcyc.org/). This is a collaborative project to build a broad network of plant metabolic pathway databases. PlantCyc, that incorporates some data from MetaCyc, is the central feature of PMN and is a database containing manually curated or reviewed information about shared metabolic pathways present in more than 300 plant species. PlantCyc serves as a reference database, while PMN also contains single species/taxon based databases. Additionally, PMN has a small number of pathways that are known to be present in other organisms and are predicted to exist in plants.

Gene Regulatory Networks

'Gene regulatory networks' consist of transcription factors and the genes that they regulate. These networks comprise of protein–DNA interactions and may also include sRNA/miRNA and sRNA/ miRNA target gene regulation. A regulatory network is formed by a series of events where regulation of one gene leads to the control of another. An example of a regulatory network database is the Arabidopsis Gene Regulatory Information Server (AGRIS) [16] which contains information on the transcription factors and cis-regulatory elements that are regulated by them in A. thaliana. AGRIS presently consists of three databases: AtcisDB, AtTFDB and AtRegNet. AtcisDB contains upstream regions of annotated A. thaliana genes and describes the experimentally validated and predicted cis-regulatory elements. AtTFDB

holds information on the transcription factors grouped into 50 conserved domain families. AtRegNet describes direct interactions between transcription factors and target genes. AGRIS also contains a Regulatory Networks Interaction Module (ReIN), that allows creation, visualization and identification of regulatory networks in A. thaliana. While AGRIS contains data from sequence annotations, TRANSFAC [17] is a gene regulatory network database that contains data on transcription factors, their experimentally proven binding sites and the genes they regulate in 300 species. TRANSFAC is one of the few proprietary plant database resources in PathGuide.

PlantCARE [18] is a database of plant cis-acting regulatory elements where the data on the transcription sites are extracted from literature supplemented with predicted data. PlantCARE provides levels of confidence for experimental evidence, functional information and position of the promoter. Additionally, a plant DNA query sequence can be searched for cis-regulatory elements using a query tool in PlantCARE.

PlantTFDB [19] is a recently constructed database that contains transcription factors from 49 plant species, grouped into 58 families. Each transcription factor is comprehensively annotated with respect to functional domains, 3D structures, gene ontology, gene expression information from expressed sequence tags (ESTs) and microarrays and annotations from other databases.

AthaMap [20] is a genome-wide map of published or experimentally determined transcription factor binding sites (TFBS) in A. thaliana. It also includes predicted sites. AthaMap allows searching for a genomic sequence or a gene to display the potential TFBS. It also provides search functionality for user defined potential co-localization

elements. Genes of interest can be analyzed for identification of common TFBSs. Conversely, genes that harbor specific TFBS can also be identified using AthaMap.

Gene co-expression network databases for plants are under development. Such databases contain information on co-expression of genes after examining a large number of experimental conditions. These can be used for identification of genes involved in a certain function, identification of cis-regulatory elements, construction of regulatory networks (although co-expression does not necessarily mean co-regulation [21]) and assist in many other biological problems. Some examples of gene co-expression networks and their applications are discussed in the Supplementary Data.

Protein–protein Interaction Networks

'Protein–protein interaction pathways' contain all interactions, stable or transient, between same or different proteins that are important for the functioning of a cell. Protein–protein interactions take place during protein modification, protein transport, protein oligomerization for activity/non-activity, chaperone assisted protein folding, signal transduction, etc. Protein–protein interaction pathways contain information on all these interactions. The A. thaliana protein interactome database (AtPID) is one such database [22]. It contains protein interaction pairs found through manual text mining or in silico predictions using various bioinformatics methods, along with protein pairs that have been confirmed.

It is now recognized that the experiments required to generate protein interaction data (e.g. yeast-twohybrid systems) often give false positives as well as false negatives and hence it is important to use this type of data with caution. To discern whether a

certain result is reliable, one needs to know the type of experiment and the conditions used, as well as details about the results. A rational assessment as to whether an interaction is truly possible in vivo can be made based on a variety of factors, including the domains involved in interaction and the type of interaction. The IntAct database [23], which contains protein–protein interaction information on several organisms including plant systems, includes such high level details.

Another database, the Predicted Arabidopsis Interactome Resource (PAIR)[24], predicts the potential interactions in A. thaliana using a support vector machine (SVM) model (a machine learning approach) and careful preparation of example data, selection of indirect evidence and a tight control of false positives. We believe that the PAIR database is currently the most accurate and comprehensive database on A. thaliana protein–protein interactions.

Combining interaction data generated through experimental and predictive methods increases the coverage of an interactome and can lead to more reliable information. When the same data is obtained through different methods one can reasonably expect more accurate data. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [25] is a multi-organism (not limited to the kingdom Plantae) database that includes all available protein–protein interactions. It scores and weighs this information and augments it with predicted interactions and automated text-mining results. STRING includes both physical and functional information on the interactions. This adds an extra measure of reliability to the interaction data.

Signaling Pathways

'Signaling pathways' comprise of molecular networks in the signal transduction cascade. These are involved in transmission of information from one part of the cell to another (intracellular, e.g. from the cytoplasm to the nucleus) or from one cell to another (intercellular, e.g. from one neuron to another). Extracellular stimuli can also bring about the activation or inhibition of a pathway and thus a change in the cellular environment. Signaling pathways often involve protein–protein interactions at different levels like protein modification (e.g. protein phosphorylation), protein translocation and protein complex formation or dissociation. Several signaling pathway databases, for example SPIKE [26], exist for non-plant eukaryotes. INOH (hosted at http://www.inoh.org/) is a signaling pathway database for Drosophila melanogaster. SignaLink (hosted at http://signalink.org/) is a cross-species database that includes pathways from human, D. melanogaster and Caenorhabditis elegans. In contrast, few plant signaling pathway databases exist and they lack the quality and efficiency in comparison to their non-plant counterparts. The DRASTIC [27] database resource for analysis of signal transduction in cells developed by the Scottish Crop Research Institute (SCRI) was one of the first relational databases in this area. It included ESTs and regulated genes in response to various environmental factors like pathogens, chemical exposure, drought, salt and low temperature. The data was collected from refereed journals. However, this reference resource is no longer available.

Recently, a database containing the stress response transcription factor database, STIFDB [28], has been created for A. thaliana. It contains the abiotic stress response genes that were found upregulated in microarray experiments, with options to identify

possible transcription factor binding sites. PathoPlant [29, 30] is another relational database that contains components of signal transduction pathways related to plant pathogenesis. It also contains microarray data of genes expressed in response to pathogens.

There is a glaring need for plant signaling pathway databases that contain and regularly update all proven and potential/putative signaling pathways in plants as these are discovered. MAPK signaling cascades were discovered >15 years ago in plants [31]. Analogues of pathways that were only known in animals are now being found as well. For example, glutamate receptors (iGluRs) that are involved in excitatory neurotransmission pathways have been extensively studied in the animal kingdom and have been included in several pathway databases. Glutamate receptor-like proteins (GLRs) were reported in 1998 in A. thaliana [32]. Since then these proteins in A. thaliana and other plants have been suggested to be involved in a wide array of pathways, through transgenic plant studies or pharmacological studies. Suggested functions include Ca2þ allocation [33], carbon/nitrogen sensing [34], regulation of abscisic acid and water balance [35], coordinating mitosis in root apical meristem [36], light signal transduction [37] and resistance to fungal infections [38]. Both MAPKs and glutamate-like receptors from A. thaliana are included in a few plant pathway databases like AtPID. However, it is difficult for a biologist looking for pathways involved in resistance to fungal infections, for example, to come immediately across the glutamate receptor-like system or conversely to find all the plant pathways that glutamate receptor like-proteins are involved in by using a keyword. Such databases would be essential to

'de-specialize' information and make it available to a wider range of scientists. This also highlights the need for such databases to be freely available to allow biologists irrespective of the system/field that they work with (plant, animal, microbial and so on) with an interest in a particular pathway to retrieve all the relevant information available.

Signaling pathway mechanisms like sugar signaling [39], light signaling [40], jasmonate signaling [41] and their components have been discovered in plants and call for dedicated pathway databases. Looking at the signaling pathways and the properties that these affect in plants, it can be concluded that these pathways cross-connect. It is important to understand these pathways and to integrate this information with other databases in order to obtain a more complete picture which would then enable plant scientists to modulate certain plant properties without affecting other mechanisms and pathways.

Pathway visualization tools

Visualization of pathway data is important not only to understand the data, but also to analyze and to build valid hypotheses based on these data. To address these requirements, many pathway/network visualization tools have been constructed with different functionalities. The level of visualization that these tools offer range from simple two-dimensional pathway maps like those provide by KEGG, to three-dimensional and hierarchical visualizations in immersive virtual reality (C6) environments like those provided by MetNetGE [42]. Interactive visualization allows users to analyze, edit and modify the pathways based on their own experimental data, as

is provided by GenMAPP [43]. Gehlenborg et al. [44] in their recent review have thoroughly reviewed available pathway visualization tools and have broadly divided these tools into two partly overlapping categories—tools focused on automated methods for interpreting and exploring large biological networks and tools focused on assembly and curation of pathways. Many of these tools integrate with public databases, allowing the users to analyze and visualize their own data. Another exhaustive overview of visualization tools has been presented by Suderman and Hallett [45]. For a critical evaluation of the requirements for biological visualization tools based on interviews conducted to understand the needs for pathway analysis, see ref. [46].

Pathway database evolution through integration

An individual pathway database holds a variety of information. This has proved to be challenging for scientists who want to access and use this information. Information is scattered across various databases that differ not only in the type of data they contain, but also the form in which they exist. Additionally, in an actual living cell, the pathways are vastly interconnected. Integration of pathway databases thus becomes imperative in order to understand a biological mechanism in its entirety. Researchers interested in a particular biological mechanism should be able to easily find and access all the data they need, without having to go through the difficult process of shifting data from different databases that are based on different platforms.

One of the biggest challenges to the integration of databases is their diversity. The existing databases have syntactic differences in the form of data file formats and

retrieval methods and semantic differences in the terminologies and data models [47]. Several pathway database resources listed in Pathguide are not machine-readable. Machine-readability is an essential requirement for automatic data retrieval and processing. Recognition of these challenges has demanded increased efforts to establish pathway ontology standards for defining models. Systems Biology Markup Language (SBML) has presented itself as one such standard for storing and sharing of computational models of biological networks [48]. Another, named BioPAX [49] was developed for detailed pathway depiction and for permitting data exchange as used in the development of MetNet [50]. PSI-MI [51] allows data exchange for protein–protein interactions, while CellML [52] enables storage and exchange of computer based mathematical models. Other data exchange formats exist that are peripherally associated with network-data and can certainly serve as input for other software packages that determine such networks. The Chemical Markup Language (CML) can be used to describe small molecules and ligands that participate in networks [53], whereas the Protein Markup Language (ProML), along with its predecessor PDB, can be used to characterize larger binding-partners [54]. The Microarray Gene Expression Markup Language (MAGE-ML) can be used as input to determine gene co-expression networks under various conditions [55]. The Ondex eXchange Language (OXL) format claims superiority over a range of formats [56], but is more general and requires more coding to implement correctly. Finally, an Application Programming Interface (API) can be provided [57], but then each API requires some study of its peculiarities (as it applies to only one particular database) as well.

Providing an easy-to-use interface for end-users is challenging with formats that allow too many options. All standards are now being used by at least some pathway databases and are certainly steps in the right direction. While laudable efforts in their own right, the proliferation of different data formats creates its own problems: providers need to decide which formats to support and each format represents a laborious and resource-intensive effort. Therefore, many times data formats still need to be converted from one format into another [58].

Ongoing efforts to automate data access and retrieval make the process much simpler for a biologist. KEGG [59] is a comprehensive resource for metabolic pathways and contained data that were originally curated manually from literature and the pathways existed as simple drawings. All pathway maps in KEGG have been redrawn, using KegSketch. The resulting KGMLþ files [60] are machine readable and editable.

Plant pathway database integration is a challenge as far fewer plant genomes have been sequenced compared to other life forms (which makes it more difficult to base inferences on homology) and the data resources on plant pathways are more dispersed [61]. The uniqueness of secondary metabolism that exists in many systems adds another layer of complexity. It is, therefore, even more important for plant pathway databases to start incorporating and supporting already existing standard formats for better integration of information and knowledge extraction. The positive side of having a limited number of plant pathway databases is that standardization needs to be applied to a smaller number of pathways. This entails less work than what would be required in other settings.

As can be seen from Figures 1 and 2 and Table 1, plant databases are still far from being overwhelmed with information and diversity load. This makes their standardization and implementation efforts much more realistic than for other systems. Furthermore, this in itself can pave the way for other systems to follow suit by learning from the successes and challenges of plant pathway database integration projects. It would therefore be a tremendously useful exercise for all upcoming plant pathway databases to start following universal standardization right from their conception. Perhaps journals should only accept the publication of databases that conform to—what we term as—Good Databasing Practice (GDbP) standards (Table 2), thereby forcing these to become standard practice. Such practices have already been incorporated for microarray and sequencing results.

Good databasing practice	Usefulness
Easy user access	Easy access for even the non-specialists
Integrated visualization tools	Ease understanding and analysis of large data sets
Standard ontology	Ease of data exchange
Possibility to integrate data from other databases	Expansion of available information
Proper documentation of stored data; provision of source and reliability of original data	Possibility to get back to the original source if required, enable judgment of accuracy of inferred information
Provision of risk factors and probability of error propagation when deriving orthologs in another species	Using particular data with caution when inferring a pathway or an ortholog
Good user support	Good response time to user queries
Regular update/maintenance	Update information; removal of errors, bugs
Regular and professional data curation and annotation	Manual curation of the data/annotations to remove errors generated by automatic data retrieval; annotation—both derived from source and inferred—help describing an entity or an event

Table 2: *Overview of Good Databasing Practices*

Applications of pathway database integration

Pathway database integration yields many potential advantages for the biologist and software developer alike. If successful, numerous applications will follow, many of

which will be surprising or even unthinkable today. To better appreciate the potential of integration, a few case studies from other fields are presented.

One study [62] integrated data from three metabolic pathways—fatty acid synthesis genes fromArabidopsis Lipid Gene Database [63] (http://lipids.plantbiology.msu.edu/), starch metabolism genes from Starch Metabolism Network project (http://www.starchmetnet.org/) and the original references for leucine catabolism—with transcriptomics data, leading to a picture that no individual study was able to show by itself. The integration revealed that each of these pathways is structured as a co-expressed module with the possibility that these modules exist in a hierarchical organization. The transcripts from each module co-accumulate over a wide range of environmental and genetic perturbations and developmental stages.

In another case study [61], A. thaliana pathways from protein interaction databases were integrated with co-expression data using the Ondex system (http://www.ondex.org/). This method enabled the determination of co-expression of the interacting protein partners and the levels of expression.

An interesting example of using database integration to obtain enhanced information about a system is AraGEM [64]. AraGEM is an attempt at building genome scale reconstruction of the primary metabolic network in A. thaliana. It used A. thaliana metabolic genome information from KEGG as a core enriched with information on the cellular compartmentalization of metabolic pathways from literature and, apart from others, databases like AraPerox [65] and Arabidopsis information resource TAIR [66]. A total of 75 essential primary metabolism reactions were identified for which genetic

information was unknown. The resulting genome-scale model was then used to construct a metabolic flux model of plant metabolism representing both photosynthetic and non-photosynthetic cell types. The model was validated by simulation of plant metabolic functions inferred from literature. AraGEM exemplifies how genome-scale models can be first built and then used to explore highly complex and compartmentalized eukaryotic networks and to construct and examine testable, non-trivial hypotheses.

A thorough literature search on plant pathways and newly discovered mechanisms can enable design of new applications through database integration. In plants, for example, hormonal and defense signaling pathways have been found to cross-talk through identical components [67]. An integration of these two types of information can point towards new targets to counteract the microbial components that decrease plant resistance and lead to disease.

Non-plant references and opportunities for the future

Human databases have already benefitted from integration of information from different pathway databases. For example, a meta-analysis study of Type-2 diabetes was conducted to find different genes that are involved in the disease. Various types of data were used: medical reviews, phenotype information, proteome analysis results, candidate gene lists from previous studies, differential gene expression and time series microarray studies [68]. The study also incorporated information from several pathway databases including KEGG, Reactome [69], BioCyc [70], GO [71], IntAct and TRANSFAC to add pathway information and to derive cellular network information on these genes. This

allowed identification of 213 genes with overall disease relevance indicating common, tissue-independent processes related to the disease and also identified genes showing changes with respect to a single study.

In another study [72], an integrated human interactome network was constructed using physical and direct binary protein–protein interactions. Data were retrieved from a variety of sources: Biomolecular Interaction Database (BIND), BioGRID, DIP, GeneRIG, IntAct, MINT and Reactome. All of these play a particular role in the integration scheme. BIND [73] contains data from large-scale cell mapping studies and molecular interactions in PDB. BioGRID [74] has protein and genetic interaction information as well as information from primary literature. DIP [75] contains experimentally determined protein–protein interactions. Gene reference into function (GeneRIF) [76] contains short text about curated articles that are relevant to known genes. IntAct contains highly curated interaction data from literature or direct deposition by experienced curators. MINT [77] focuses on experimentally verified protein–protein interactions and Reactome is a knowledgebase containing interaction data in different pathways. The Hepatitis C virus (HCV)-host infection network that was generated experimentally and from text mining was also incorporated on top of this integrated interactome network——a type of meta-integration. This led to the identification of previously unknown, novel functional pathways of HCV biology and its pathogenesis. One could extrapolate the advantages of a similar approach followed for crop plant systems and pathogens that could then divulge information on plant host–pathogen interactions and the pathways involved in pathogenesis. This could lead to development

of methods to bestow pathogen resistance on crop plants or target these pathways against the pathogen.

Not only can plant science benefit from the animal pathway database and integration examples, animal biologists can in turn benefit from the study of plant pathways by asking the question whether pathways discovered only in plants to date also exist in animals or how similar or different are the pathway networks that exist both in plants and animals. Many opportunities become available through such a feedback loop: can we unlock more evolutionary secrets? Can we become better at harnessing plants for our use or could human diseases be experimentally modeled in plants if common pathways do indeed exist for plants and animals? Applications are endless and the potential for knowledge creation extreme.

A survey of integrated pathway databases and tools

Two approaches exist to perform database integration: through the use of tools and through already integrated databases [78] (that hopefully get rebuilt periodically to stay current). Pathway database integration tools along with integrated pathway databases play a very important role in easing data integration for biologists. These tools can also be used for various other purposes like data visualization, pathway prediction, pathway gap-fillers and biological network analysis. Applications of pathway databases and tools help further knowledge of the pathways and on the inner workings of living systems.

Pathway database tools for plant systems are important because of the widely dispersed information within several databases and a lack of consistency among these databases. A growing need exists to bring this information together in a standard format to aid access and model-building. Plants show more heterogeneity among different species (e.g. in terms of secondary metabolism [79]). This makes it even more important to integrate pathway data for all important plant species and to design tools that would aid in pointing out interspecies similarities and differences.

A separate version of Reactome, Arabidopsis Reactome, [80] represents a knowledgebase of biological processes in A. thaliana and several other plant species. It integrates pathway information curated in-house, as well as from KEGG and AraCyc. It also provides a platform to navigate and discover interconnected pathways in A. thaliana. The data model of Arabidopsis Reactome uses reactions and their interconnections; it treats protein modifications, proteins localized in different compartments, as well as protein complexes, as entities on their own. It furthermore allows generalization of protein isoforms, paralogues and splice variants with a possibility of tracing these components back. The model contains both real and inferred data along with proper annotations that allow distinction between the two.

Tools like CORNET [81] help integrate A. thaliana related microarray expression data. The data sets for CORNET were obtained from Gene Expression Omnibus (GEO) [82] and from experiments carried out on Affymetrix ATH1 arrays. Also retrieved were the corresponding meta-data (which is unstructured and hence cumbersome to retrieve and parse automatically), including information about sample tissues, treatments and

sampling time points, protein interaction data, localization data and functional information. The meta-data have manually assigned ontology terms using Plant ontology [83–85], the Microarray gene expression data (MGED) ontology (MO) [86] and the Plant environmental ontology (EO) (www.gramene.org/plant_ontology/index.html#eo). Protein–protein interactions were obtained from BIND, IntAct, BioGRID, DIP, MINT, TAIR. Predicted PPIs were obtained from the BAR Arabidopsis interaction viewer [87] and AtPID. Information was also obtained from their own study [88]. Localization data were obtained from SUBA [89], iPSORT [90], LOCtree [91], MITOPRED [92], MitoProt [93], MultiLoc [94], PeroxiP [95], Predotar [96], SubLoc [97], TargetP [98] and WoLF_PSORT [99]. CORNET includes all available data along with related meta-data. The tool then provides a reliability score for each result based on the search options, parameters and thresholds used (supplied by the user). A visualization tool additionally allows the users to distinguish more reliable predictions from less predictable ones.

CORNET aims to provide functional context to genes and conversely, to provide an ability to predict functions of genes that have unknown functions. It is a tool that could also, in the future, use the information on A. thaliana to extrapolate networks in other plant species.

Many pathway resources use only the general localization predictors. In contrast, CORNET has made an attempt to also use species-specific localization information. Thus, CORNET uses localization data from both ‘general’ localization predictors and from an A. thaliana specific localization database SUBA, which was the only species-

specific resource available then. SUBA contains data retrieved from literature, experiments and from prediction tools. It has become clearer over time that use of organism-specific predictors and multiple (general) predictors are likely to lead to more accurate predicted localization [100–103]. Predictions from general predictors may not be suitable for predicting localization of an individual organism as these prediction tools are trained on proteins from a variety of organisms (and can suffer from sampling bias). Localization data from any single predictor needs to be treated with caution keeping in mind that inclusion of false positives into the integrated databases would result in amplification of the wrong information. Fortunately for plants, some organism-specific localization predictors have recently become available, e.g. AtSubP (Arabidopsis)[103] and RSLpred (rice) [104]. These should be used while integrating pathway information for the respective species. If a tool similar to CORNET is developed for rice, RSLpred would definitely be an important resource for protein localization data. A need for localization predictors specific to a variety of plants cannot be emphasized enough for a more reliable extrapolation of networks.

The ‘MetNet’ platform contains both metabolic and regulatory networks of A. thaliana, soybean [50] and grapevine. It is an attempt to integrate metabolic data from AraCyc and regulatory data from AGRIS, with additional manually curated signal transduction pathways (in A. thaliana). The pathway information is integrated with other resources like TAIR, GO-classifications (retrieved through TAIR) and MapMan [105] that supply gene related information. Protein information is obtained from PPDB [106], AMPDB [107], AtNoPDB [108], AraPerox, PLprot [109], SUBA and BRENDA [110].

These also provide the subcellular localization information for the entities. Metabolite data from ChEBI [111], PubChem [112], KEGG, NCI [113] are also integrated into the database. As there are large holes in the information on the function of a large number of genes in A. thaliana, MetNet is aimed at formulating testable hypotheses. MetNet supports various types of users and data retrieval methods. MetNet Online (available at http://metnetonline.org/) is an online interface to MetNet. MetNetAPI is an Application Programming Interface to the platform that facilitates automated data retrieval [57] and a plug-in exists for the CellDesigner environment [114].

'VitisNet' [115] is a web-based tool for grapevine (Vitis vinifera) that integrates metabolomic, proteomic and transcriptomic pathway information within molecular networks like metabolic or signaling networks and presents a molecular network model. VitisNet allows visualization of genes and biochemical pathways involved in growth, fruiting cycles and environmental stress response. Data from VitisNet is now also available in MetNet.

'Metacrop' [116] contains manually curated metabolic pathway information in crop plants (with special emphasis on seeds and tubers), along with a wide variety of other factors like reactions, location, transport processes, kinetics, taxonomy and literature. MetaCrop has an easy to use web interface and allows automatic export of information for creation of metabolic models.

Pathway database maintenance – an easily overlooked detail

Although Pathguide lists more than 300 pathway resources, at least 30 of these databases and resources are no longer functional. At the time of writing this review (October 2010), inaccessible databases 'not' marked as non-functional in Pathguide include aMAZE [117,118], Sentra [119] and EMP [120] among others. Other databases may change location. During the preparation of this article, this happened with AtPID. The publication on AtPID is now destined to refer to an incorrect URL. Several of these databases contained high quality data and unavailability of databases is a loss from several angles. For example, aMAZE boasted an excellent data model. It could deal with metabolic, protein–protein interaction, gene regulation, sub-cellular localization, signal transduction and transport and thus had the capacity to integrate a large variety of data. Its current absence is a significant loss to the scientific community at large. While papers do exist for many of these projects, the technical details of an implementation can often only be obtained through communication with the implementing team. This effectively means that if anyone else ventures to do the same elsewhere in the world, they will have to retrace the time and steps to achieve the quality of aMAZE. Similarly, Arabidopsis Reactome is another dedicated database on A. thaliana, which is currently no longer being developed as the continuation of this project requires new funding initiatives.

Due to their ever expanding and evolving nature, pathway databases (like any other scientific database) need to be maintained, curated and developed on a long-term basis. Finding financial support for long-term maintenance of pathway databases is a challenging task. One possibility is to raise funds by establishing license purchase

requirements for the use of databases, but this restricts open access to the information contained therein and can thus hinder the development of the field [121]. In addition, this is unfeasible for smaller projects that attract limited attention, but may be useful as part of integration efforts. Solutions are needed to ensure provision of continued funding for especially promising databases (without promoting an uncontrolled proliferation of new platforms) and avoid the loss of valuable information in established resources. Loss of such databases is not only a loss to the scientific community, but also is a waste of resources that have been spent on the creation and development of an excellent database in the first place. Funding agencies could, for example, provide continued funding to the database projects that they have already funded provided that the projects follow the GDbP standards which are continually and rigorously monitored and reported by an independent workgroup. Another solution could be an integration of especially promising databases into more permanent structures such as Gramene or NCBI.

The Arabidopsis Information Resource (TAIR) funding can serve as a recent example of search for alternative funding sources. NSF funding for TAIR would phase out over the next 3 years (http://www.nature.com/news/2009/091118/full/462258b.html). For its continued maintenance, TAIR has recently come up with a corporate sponsorship program. The idea is to avoid subscription requirements for the corporate sector and thus keep the resource open and free of login requirements, thereby allowing continued open access to the data for all scientists. TAIR has already secured several corporate sponsors through this program. Such programs would certainly help survival of at least some databases.

However, this is not a real alternative to public funding as such a solution could end up introducing a corporate bias into the system—only those database would survive that are able to find corporate sponsorship. Various funding models for these community resources (that are not necessarily research-projects in their own right) have recently received more attention [122, 123]. These could be applied for plant pathway database integration and maintenance. Funding a community resource requires a different approach compared to more conventional research projects. Various scenarios for databases need to be discussed and changed, a recommendation also posited by Bastow and Leonelli [123].

Conclusion

Pathway databases play an important role in advancing our knowledge of the biological functions and mechanisms. Increased understanding of living systems as a whole can, in turn, aid successful application design in silico, in vitro and in vivo.

Plants are important as veritable food, drug and fuel sources, as well as bioremediation and biotechnological tools. This provides a strong incentive to create better, more integrated and easily accessible plant pathway databases. Such efforts would lead to discovery and elucidation of the yet unknown components involved in various pathways and their function. This would also result in the creation of testable models that can further enrich the knowledge on plant systems. This then could lead to the design of more specialized intervention technologies along with potential commercial applications: innovation as a result of integration.

References

1. Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. Nucleic Acids Res 2006;34:D504–6.

2. Korc M. Pathways for aberrant angiogenesis in pancreatic cancer. Mol Cancer 2003;2:8.

3. Tsiantis M, Brown MI, Skibinski G, et al. Disruption of auxin transport is associated with aberrant leaf development in maize. Plant Physiol 1999;121:1163–8.

4. Kelley BP, Sharan R, Karp RM, et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci USA 2003;100: 11394–9.

5. Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol 2000;18:609–13.

6. Bumgarner RE, Yeung KY. Methods for the inference of biological pathways and networks. Methods Mol Biol 2009; 541:225–45.

7. Pradines J, Rudolph-Owen L, Hunter J, et al. Detection of activity centers in cellular pathways using transcript profiling. J BiopharmStat 2004;14:701–21.

8. Apic G, Ignjatovic T, Boyer S, et al. Illuminating drug discovery with biological pathways. FEBS Lett 2005;579: 1872–7.

9. Schreiber SL. Target-oriented and diversity-oriented organic synthesis in drug discovery. Science 2000;287: 1964–9.

10. Dixon N, Duncan JN, Geerlings T, et al. Reengineering orthogonally selective riboswitches. Proc Natl Acad Sci USA 2010;107:2830–5.

11. Mueller LA, Zhang P, Rhee SY. AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiol 2003;132: 453–60.

12. Huang YJ, Hang D, Lu LJ, et al. Targeting the human cancer pathway protein interaction network by structural genomics. Mol Cell Proteomics 2008;7:2048–60.

13. Lynn DJ, Winsor GL, Chan C, et al. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Mol Syst Biol 2008;4:218.

14. Caspi R, Foerster H, Fulcher CA, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 2008;36:D623–31.

15. Liang C, Jaiswal P, Hebbard C, et al. Gramene: a growing plant comparative genomics resource. Nucleic Acids Res 2008;36:D947–53.

16. Davuluri RV, Sun H, Palaniswamy SK, et al. AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics 2003;4:25.

17. Matys V, Fricke E, Geffers R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. NucleicAcidsRes 2003;31:374–8.

18. Lescot M, Dehais P, Thijs G, etal. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 2002;30:325–7.

19. Guo AY, Chen X, Gao G, et al. PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res 2008;36:D966–9.

20. Bulow L, Steffens NO, Galuschka C, et al. AthaMap: from in silico data to real transcription factor binding sites. In Silico Biol 2006;6:243–52.

21. Stuart JM, Segal E, Koller D, et al. A gene-coexpression network for global discovery of conserved genetic modules. Science 2003;302:249–55.

22. Cui J, Li P, Li G, et al. AtPID: Arabidopsis thaliana protein interactome database–an integrative platform for plant systems biology. Nucleic Acids Res 2008;36:D999–1008.

23. Aranda B, Achuthan P, Alam-Faruque Y, et al. The IntAct molecular interaction database in (2010). Nucleic Acids Res 2010;38:D525–31.

24. Lin M, Shen X, Chen X. PAIR: the predicted Arabidopsis interactome resource. Nucleic Acids Res 2011. [Epub ahead of print 2010].

25. Jensen LJ, Kuhn M, Stark M, et al. STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 2009;37:D412–6.

26. Elkon R, Vesterman R, Amit N, et al. SPIKE–a database, visualization and analysis tool of cellular signaling pathways. BMC Bioinformatics 2008;9:110.

27. Button DK, Gartland KM, Ball LD, et al. DRASTIC– INSIGHTS: querying information in a plant gene expression database. Nucleic Acids Res 2006;34:D712–6.

28. Shameer K, Ambika S, Varghese SM, et al. STIFDB-Arabidopsis Stress Responsive Transcription Factor DataBase. IntJ Plant Genomics 2009;2009:583429.

29. Bulow L, Schindler M, Choi C, et al. PathoPlant: a database on plant-pathogen interactions. In Silico Biol 2004;4: 529–36.

30. Bulow L, Schindler M, Hehl R. PathoPlant: a platform for microarray expression data to analyze co-regulated genes involved in plant defense responses. Nucleic Acids Res 2007; 35:D841–5.

31. Rodriguez MC, Petersen M, Mundy J. Mitogen-activated protein kinase signaling in plants. AnnuRev Plant Biol 2010; 61:621–49.

32. Lam HM, Chiu J, Hsieh MH, et al. Glutamate-receptor genes in plants. Nature 1998;396:125–6.

33. Kim SA, Kwak JM, Jae SK, et al. Overexpression of the AtGluR2 gene encoding an Arabidopsis homolog of mammalian glutamate receptors impairs calcium utilization and sensitivity to ionic stress in transgenic plants. Plant Cell Physiol 2001;42:74–84.

34. Kang J, Turano FJ. The putative glutamate receptor 1.1 (AtGLR1.1) functions as a regulator of carbon and nitrogen metabolism in Arabidopsis thaliana. Proc Natl Acad Sci USA 2003;100:6872–7.

35. Kang J, Mehta S, Turano FJ. The putative glutamate receptor 1.1 (AtGLR1.1) in Arabidopsis thaliana regulates abscisic acid biosynthesis and signaling to control development and water loss. Plant Cell Physiol 2004;45:1380–9.

36. Li J, Zhu S, Song X, et al. A rice glutamate receptor-like gene is critical for the division and survival of individual cells in the root apical meristem. Plant Cell 2006;18:340–9.

37. Brenner ED, Martinez-Barboza N, Clark AP, et al. Arabidopsis mutants resistant to S(þ)-beta-methyl-alpha, beta-diaminopropionic acid, a cycad-derived glutamate receptor agonist. Plant Physiol 2000;124:1615–24.

38. Kang S, Kim HB, Lee H, et al. Overexpression in Arabidopsis of a plasma membrane-targeting glutamate receptor from small radish increases glutamate-mediated Ca2þ influx and delays fungal infection. Mol Cells 2006; 21:418–27.

39. Bolouri-Moghaddam MR, Le Roy K, Xiang L, et al. Sugar signalling and antioxidant network connections in plant cells. FEBSJ 2010;277:2022–37.

40. Chory J. Light signal transduction: an infinite spectrum of possibilities. PlantJ 2010;61:982–991.

41. Chung HS, Niu Y, Browse J, et al. Top hits in contemporary JAZ: an update on jasmonate signaling. Phytochemistry 2009;70:1547–59.

42. Jia M, Choi SY, Reiners D, et al. MetNetGE: interactive views of biological networks and ontologies. BMC Bioinformatics 2010;11:469.

43. Dahlquist KD, Salomonis N, Vranizan K, et al. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 2002;31:19–20.

44. Gehlenborg N, O'Donoghue SI, Baliga NS, et al. Visualization of omics data for systems biology. Nat Methods 2010;7:S56–68.

45. Suderman M, Hallett M. Tools for visually exploring biological networks. Bioinformatics 2007;23:2651–9.

46. Saraiya P, North C, Duca K. Visualizing biological pathways: requirements analysis, systems evaluation and research agenda. InfVis 2005;4:191–205.

47. Cary MP, Bader GD, Sander C. Pathway information for systems biology. FEBS Lett 2005;579:1815–20.

48. Hucka M, Finney A, Sauro HM, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 2003;19:524–31.

49. Luciano JS. PAX of mind for pathway researchers. Drug DiscovToday 2005;10:937–42.

50. Wurtele ES, Li L, Berleant D, et al. MetNet: Systems Biology Software for Arabidopsis. In: Nikolau BJ, Wurtele ES, (eds). Concepts in Plant Metabolomics. Springer, 2007;145–58.

51. Hermjakob H, Montecchi-Palazzi L, Bader G, et al. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. Nat Biotechnol 2004;22:177–83.

52. Lloyd CM, Halstead MD, Nielsen PF. CellML: its future, present and past. Prog BiophysMol Biol 2004;85:433–50.

53. Liao YM, Ghanadan H. The chemical markup language. Anal Chem 2002;74:389A–90A.

54. Hanisch D, Zimmer R, Lengauer T. ProML–the protein markup language for specification of protein sequences, structures and families. In Silico Biol 2002;2:313–24.

55. Spellman PT, Miller M, Stewart J, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 2002;3: RESEARCH0046.

56. Taubert J, Sieren KP, Hindle M, et al. The OXL format for the exchange of integrated datasets. J Integr Bioinf 2007;4: 62–75.

57. Sucaet Y, Wurtele ES. MetNetAPI: A flexible method to access and manipulate biological network data from MetNet. BMCResNotes 2010;3:312.

58. Heath AP, Kavraki LE. Computational challenges in systems biology. Comput Sci Rev 2009;3:1–17.

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

60. Kanehisa M, Goto S, Furumichi M, et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 2010;38:D355–60.

61. Lysenko A, Hindle MM, Taubert J, et al. Data integration for plant genomics–exemplars from the integration of Arabidopsis thaliana databases. Brief Bioinform 2009;10: 676–93.

62. Mentzen WI, Peng J, Ransom N, et al. Articulation of three core metabolic processes in Arabidopsis: fatty acid biosynthesis, leucine catabolism and starch metabolism. BMCPlant Biol 2008;8:76.

63. Beisson F, Koo AJ, Ruuska S, et al. Arabidopsis genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. PlantPhysiol 2003; 132:681–97.

64. de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW, et al. AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. Plant Physiol 2010; 152:579–89.

65. Reumann S, Ma C, Lemke S, et al. AraPerox. A database ofputative Arabidopsis proteins from plant peroxisomes. Plant Physiol 2004;136:2587–608.

66. Poole RL. The TAIR database. MethodsMol Biol 2007;406: 179–212.

67. Grant MR, Jones JD. Hormone (dis)harmony moulds plant health and disease. Science 2009;324:750–752.

68. Rasche A, Al-Hasani H, Herwig R. Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus. BMCGenomics 2008; 9:310.

69. Matthews L, Gopinath G, Gillespie M, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 2009;37:D619–22.

70. Karp PD, Ouzounis CA, Moore-Kochlacs C, et al. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res 2005; 33:6083–9.

71. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9.

72. de Chassey B, Navratil V, Tafforeau L, et al. Hepatitis C virus infection protein network. Mol Syst Biol 2008;4:230.

73. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 2003;31: 248–50.

74. Stark C, Breitkreutz BJ, Reguly T, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535–9.

75. Xenarios I, Salwinski L, Duan XJ, etal. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002;30: 303–5.

76. Lu Z, Cohen KB, Hunter L. GeneRIF quality assurance as summary revision. Pac Symp Biocomput 2007;269–80.

77. Chatr-aryamontri A, Ceol A, Palazzi LM, et al. MINT: the Molecular INTeraction database. Nucleic Acids Res 2007;35: D572–4.

78. Neerincx PB, Leunissen JA. Evolution of web services in bioinformatics. Brief Bioinform 2005;6:178–188.

79. Schwab W. Metabolome diversity: too few genes, too many metabolites? Phytochemistry 2003;62:837–49.

80. Tsesmetzis N, Couchman M, Higgins J, et al. Arabidopsis reactome: a foundation knowledgebase for plant systems biology. Plant Cell 2008;20:1426–36.

81. De Bodt S, Carvajal D, Hollunder J, et al. CORNET: a user-friendly tool for data mining and integration. Plant Physiol 2010;152:1167–79.

82. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol 2006;411:352–69.

83. Bruskiewich R, Coe EH, Jaiswal P, etal. The plant ontology consortium and plant ontologies. Comp Funct Genomics 2002;3:137–42.

84. Avraham S, Tung CW, Ilic K, et al. The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Res 2008;36:D449–54.

85. Ilic K, Kellogg EA, Jaiswal P, et al. The plant structureontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant Physiol 2007;143:587–99.

86. Whetzel PL, Parkinson H, Causton HC, et al. The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics 2006;22:866–73.

87. Geisler-Lee J, O'Toole N, Ammar R, et al. A predicted interactome for Arabidopsis. Plant Physiol 2007;145:317–29.

88. De Bodt S, Proost S, Vandepoele K, et al. Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression. BMCGenomics 2009;10:288.

89. Heazlewood JL, Verboom RE, Tonti-Filippini J, et al. SUBA: the Arabidopsis Subcellular Database. Nucleic Acids Res 2007;35:D213–8.

90. Bannai H, Tamada Y, Maruyama O, et al. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 2002;18:298–305.

91. Nair R, Rost B. Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol 2005;348: 85–100.

92. Guda C, Fahy E, Subramaniam S. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. Bioinformatics 2004;20:1785–94.

93. Claros MG. MitoProt, a Macintosh application for studying mitochondrial proteins. Comput Appl Biosci 1995;11:441–7.

94. Hoglund A, Donnes P, Blum T, et al. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics 2006;22:1158–65.

95. Emanuelsson O, Elofsson A, von Heijne G, et al. In silico prediction of the peroxisomal proteome in fungi, plants and animals. JMol Biol 2003;330:443–56.

96. Small I, Peeters N, Legeai F, et al. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 2004;4:1581–90.

97. Chen H, Huang N, Sun Z. SubLoc: a server/client suite for protein subcellular location based on SOAP. Bioinformatics 2006;22:376–7.

98. Emanuelsson O, Brunak S, von Heijne G, et al. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2007;2:953–71.

99. Horton P, Park KJ, Obayashi T, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res 2007;35: W585–7.

100. Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng 1999;12:3–9.

101.Bender A, van Dooren GG, Ralph SA, et al. Properties and prediction of mitochondrial transit peptides from Plasmodium falciparum. Mol Biochem Parasitol 2003; 132:59–66.

102. Schneider G, Fechner U. Advances in the prediction of protein targeting signals. Proteomics 2004;4:1571–80.

103.Kaundal R, Saini R, Zhao PX. Combining machine learning and homology-based approaches to accurately predict subcellular localization in Arabidopsis. Plant Physiol 2010; 154:36–54.

104.Kaundal R, Raghava GP. RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. Proteomics 2009;9:2324–42.

105.Thimm O, Blasing O, Gibon Y, et al. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. PlantJ 2004;37:914–39.

106.Sun Q, Zybailov B, Majeran W, et al. PPDB, the Plant Proteomics Database at Cornell. Nucleic Acids Res 2009;37: D969–74.

107.Heazlewood JL, Tonti-Filippini JS, Gout AM, et al. Experimental analysis of the Arabidopsis mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. Plant Cell 2004;16:241–56.

108.Brown JW, Shaw PJ, Shaw P, et al. Arabidopsis nucleolar protein database (AtNoPDB). Nucleic Acids Res 2005;33: D633–6.

109. Kleffmann T, Hirsch-Hoffmann M, Gruissem W, et al. plprot: a comprehensive proteome database for different plastid types. Plant Cell Physiol 2006;47:432–6.

110.Chang A, Scheer M, Grote A, et al. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in (2009). NucleicAcidsRes 2009;37:D588–92.

111. Degtyarenko K, de Matos P, Ennis M, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 2008;36:D344–50.

112.Wang Y, Xiao J, Suzek TO, et al. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 2009;37:W623–33.

113. Sitzmann M, Filippov IV, Nicklaus MC. Internet resources integrating many small-molecule databases. SAR QSAR EnvironRes 2008;19:1–9.

114.Van Hemert JL, Dickerson JA. PathwayAccess: CellDesigner plugins for pathway databases. Bioinformatics 2010;26:2345–6.

115. Grimplet J, Cramer GR, Dickerson JA, et al. VitisNet: ‘‘Omics’’ Integration through Grapevine Molecular Networks. PLoSONE 2009;4:e8365.

116. Grafahrend-Belau E, Weise S, Koschutzki D, et al. MetaCrop: a detailed database of crop plant metabolism. Nucleic Acids Res 2008;36:D954–8.

117.Lemer C, Antezana E, Couche F, et al. The aMAZE LightBench: a web interface to a relational database of cellular processes. Nucleic Acids Res 2004;32:D443–8.

118. van Helden J, Naim A, Lemer C, et al. From molecular activities and processes to biological function. Brief Bioinform 2001;2:81–93.

119.D’Souza M, Glass EM, Syed MH, et al. Sentra: a database of signal transduction proteins for comparative genome analysis. Nucleic Acids Res 2007;35:D271–3.

120. Selkov E, Basmanova S, Gaasterland T, et al. The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. Nucleic Acids Res 1996;24:26–8.

121. Philippi S, Kohler J. Addressing the problems with life-science databases for traditional uses and systems biology. Nat Rev Genet 2006;7:482–8.

122. Chandras C, Weaver T, Zouberakis M, et al. Models for financial sustainability of biological databases and resources. Database 2009;106:17475–80.

123.Bastow R, Leonelli S. Sustainable digital infrastructure. EMBORep 2010;11:730–4.

CHAPTER II

METNET ONLINE

A Novel Integrated Resource For Plant Systems Biology

Plants are important as drug resources, biofuel resources, bioremediation tools, and general tools for recombinant technology. Study of plant biological pathways requires easy access to (preferably already integrated) data sources. Today, various plant data sources are scattered throughout the web, making it hard to build comprehensive datasets.

MetNet Online is a web-based portal that provides access to a regulatory and metabolic plant pathway database. The database and portal integrate Arabidopsis, Soybean (Glycine max), and Grapevine (Vitis vinifera) data. Pathways are enriched with known or predicted information on subcellular location. MetNet Online enables pathways, interactions, and entities to be browsed or searched by multiple categories, such as subcellular compartment, pathway ontology, and GO term. In addition, the “My MetNet” feature allows registered users to bookmark content and track, import, and export customized lists of entities. Users can also construct custom networks using existing pathways and/or interactions as building blocks.

The site can be reached at http://www.metnetonline.org. Extensive video-tutorials on how to use the site are available through http://www.metnetonline.org/tutorial/.

Introduction and background

Plants are increasingly facilitating and augmenting (quality of) human life, and plant systems biology resources exist in a variety of locations [1]. Researchers interested in a particular biological mechanism should be able to easily find and access all the data they need, without having to go through the difficult process of shifting data from different databases that are based on different platforms. This provides a strong incentive to create better, more integrated and easily accessible integrated plant portals.

A biological network database needs to capture and represent biological relationships in many ways. MetNet consists of a suite of software tools that specialize in different areas of systems biology [2-10]. Our database currently contains information about three model plants (Arabidopsis, soybean and grapevine). An exhaustive list of integrated resources is available online. In additional to the retrieved information, manual curation took place. Regulatory information is one type of data that was added manually with input from expert biologists. MetNet Online provides an easy-to-use front-end web interface and combines several important features to provide an unique platform. First, metabolism, signalling, and transcriptional pathways are fully integrated into a single network. Second, a subcellular location layer (obtained via manual curation and/or information from extant databases) overlays the pathways. Third, a protein-protein interaction layer extends pathway information. Fourth, the website allows for customized views of any data: any combination of pathways and interactions can be combined into a new network). Fifth, MetNet Online has a “My MetNet” component, which operates similarly to “My NCBI”. Users can keep track of (bookmark) their

favorite entities, as well as lists of particular interest (e.g., lists of genes upregulated in a given mutant, or metabolites derived from cytosolic acetyl-CoA). Lastly, the search-function is sufficiently intelligent to recognize synonyms (e.g., "water" is listed amongst the search results whether one searches for "H2O" or "water"). MetNet Online is complementary to other community resources and provides a starting point for researchers to develop new hypotheses about biological function [11]. To enable the user to easily analyze network data and customized content in MetNet Online, we provide different ways to export data (including Graphviz .dot, SBML, and XGMML) to facilitate data-flow to external applications. For bioinformatics software developers, a separate application programming interface (API) is provided [12].

Results and implementation

The MetNet Online portal

MetNet Online is a web application developed in PHP [13] using MySQL (http://www.mysql.com) as a back-end database. GraphViz (http://www.graphviz.org) is used to generate graphical representations of pathways and networks (Figure 3). Our network database is stored as an integrated labeled graph model and represents complex internal relationships. Biological entities and interactions are represented as nodes and the associations between them are represented as edges in the graph model. The database serves as the primary data repository for both our online portal and the MetNet suite of visualization and analysis tools [7].

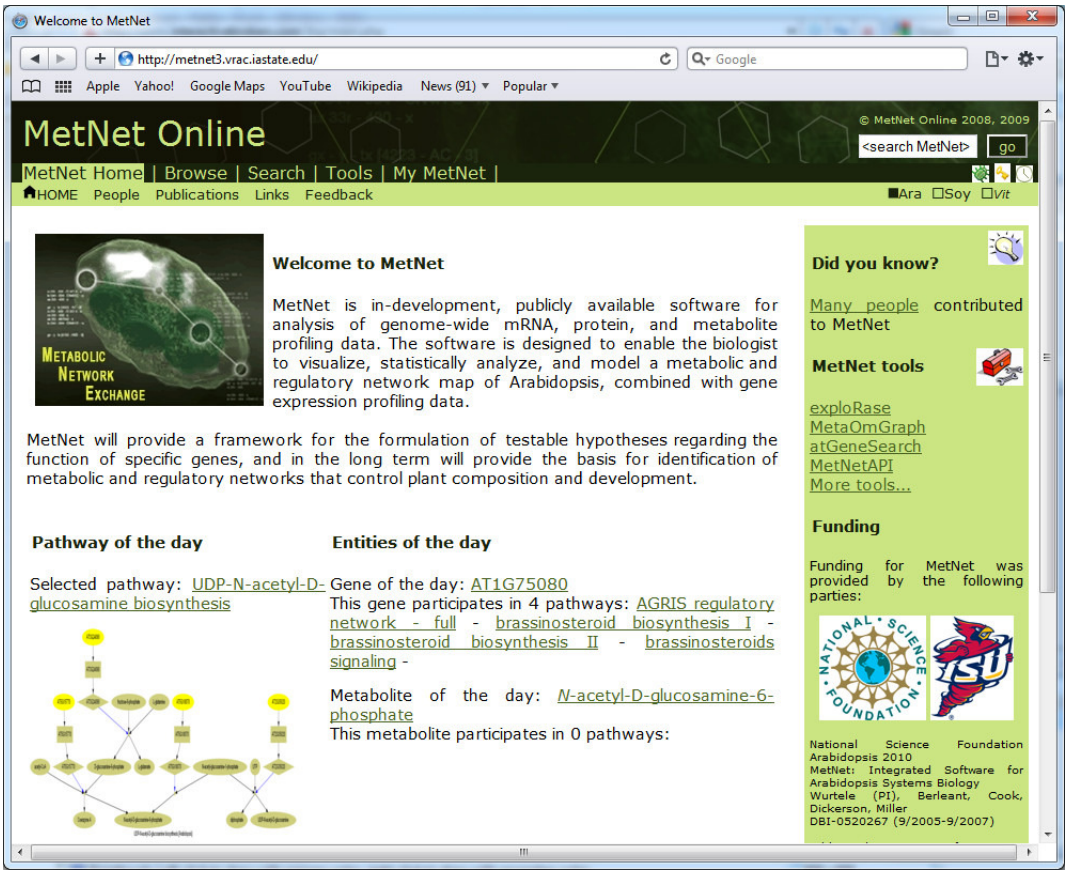

Figure 3: *The MetNet Online portal start page. In order to rapidly familiarize novice users, a "pathway of the day" display and a "gene of day" display encourage self-guided exploration.*

MetNet Online is centered on several concepts that are inherent in the underlying database. Entities represent physical molecules (subtypes are DNA, RNA, protein, protein complex, and metabolite). Interactions can occur between any number of entities or between entities and other interactions (e.g., in the case of catalysis). A pathway is a collection of interactions. Pathways are predefined and cannot be changed by a user. A network is a collection of interactions for which the granularity is determined by the user when (s)he creates it. A network can consist of any number of interactions or it can be a combination of some already defined pathways. It can also map to exactly one pathway

or it can map to a pathway minus transcription/translation events. Networks are virtual and transient objects.

The database can be browsed based on different ontologies or navigation trees including pathway category (e.g., biosynthesis, respiration, and signal transduction), entity participation, cellular location (e.g., nucleus, plastid, and cytosol) and interaction type (e.g., diffusion, transport, and negative/positive regulation). After navigating through a tree and selecting a node of interest, a list of pathways is displayed (either in list-form or by thumbnail). The pathway information screen can then be chosen or the pathway can be visualized directly.

Information about a pathway consists of general comments and literature references, location information, interactions contained within the pathway, and participating entities. Sources and sinks for the pathway are displayed in a separate tab as part of the participating entities. This is critical information for simulations in which the pathway is treated as a black box (e.g., for the glycolysis-pathway, glucose would be a source and pyruvate would be a sink; ATP and ADP would serve as both source and sink). At the top-right of the pathway information screen, a toolbar is shown with export-functions to various programs and a link to visualize the pathway (discussed separately). Entities can be browsed (alphabetically) independently of pathways. An entity information screen contains location information, possible synonyms, pathway participation, and categorized interactions. Additional tabs are available in the pathway information screen. The literature tab interfaces with PubMed to retrieve a current literature feed, whereby the name of the pathway is used as a search-term. Cellular

context (i.e. the location within the cell where an entity is present when participating in an interaction) is represented separately, so that one can get an idea of the various roles a protein or metabolite might play. Another tab shows connected pathways that share one or more entities.

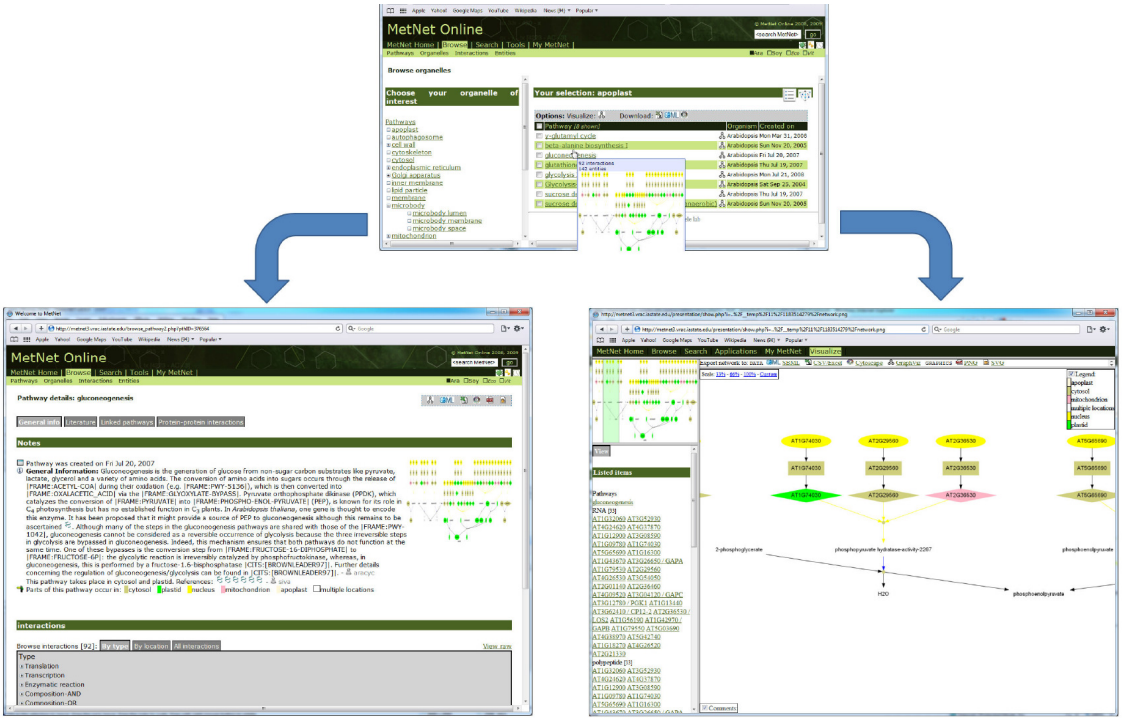

Figure 4: *Different browsing options. Pathways can be browsed by the subcellular location where (part of) the pathway occurs. Hovering over a pathway in the right-side panel bring up a thumbnail of the pathway. The pathway can then be browsed in textual mode (enumerated list of interactions and entities that make up the pathway) and visual mode.*

MetNet Online visualizes pathways with their known or predicted subcellular locations (Figure 4). This is information that is not available anywhere else: out of 5527 proteins in AraCyc 8.0 e.g., only 286 have a location annotation. Subcellular location information can help scientists develop hypotheses on gene function. Entities are color-

coded according to assigned location, and shape-coded according to entity type. Interactions are color-coded according to type for easy identification and visualization.

MetNet Online's search function is integrated, rather than providing different search functions for entities, interactions, and pathways. Thus, search operations for "regulation", "biosynthesis", "AT4G40090", "AGP3", or "malate" use the same interface. Search-results are grouped by entity types, interactions, and pathways. When searching for "glucose", not only the "glucose" metabolite is presented but also the "glucose-UDP biosynthesis" pathway, among others. Synonyms are taken into account: A search for "H2O" and "water" or "O2" and "oxygen" both lead to the same entity. Typos and misspellings result in suggestions that often point a visitor in the right direction. When no results are found for a search, potential alternatives are suggested. When "giberelin" is entered, the alternative "gibberellin" is proposed.

When visualizing a pathway, a GraphViz (http://www.graphviz.org) .dot file is generated and transformed into its visual representation (dot layout). In the upper-left corner of the screen a thumbnail of the complete pathway is shown to allow easy navigation in complex maps. An indexed list of all participating entities is displayed underneath the thumbnail.

Custom network design and personalization features

Most pathway databases are static; With MetNet Online, all lists of interactions and pathways are represented with checkboxes in front of them. A user can then choose either to visualize a pathway by itself, or to select a set of pathways or interactions. This generates a new network, representing an integrated view of all selections. The resulting

network is visualized using the methods discussed earlier or it can be exported as a whole to other tools.

Any visitor can become a registered user of MetNet Online. This opens up access to the “My MetNet” function, which is implemented in a format that is similar to other personalization portals such as “My NCBI” or “My Yahoo”. When logged in, users gain access to additional functionality. Bookmarks are used to easily retrieve objects of interest at any time in the future without having to navigate classification trees or execute a search first.

Entities, interactions, and pathways can all be bookmarked. Bookmarked objects can have commentary attached to them. “List of entities” is a second function in “My MetNet”. Users can multiple lists simultaneously, which can be created in three ways: a user can manually specify its members, convert a set of bookmarked entities into a list, or upload a text-file. An entity list can include experimental data, such as over-represented or under-represented genes from transcriptomics analysis or metabolites from a GC/MS experiment. A list of genes can be forwarded to Reactome’s Skypainter function [14, 15].

As a list gets longer, it is likely that additional pathways will be linked to the entities in that list. In order to put results in perspective and to distinguish relevant from less relevant pathways, a separate interface is presented that contains the results of Fisher’s exact test and which ranks matching pathways by p-value (lesser values indicate higher relevance). Fisher’s exact test is available for both visitors and registered users but registered users can automatically apply the test to lists of entities that have been

created previously. Visitors will need to specify their entities of interest manually in a text field.

By providing the option to export pathways to other file formats, MetNet Online leverages existing software that incorporates a range of supplementary layout algorithms (CellDesigner[42], Cytoscape [43]) in a more suitable environment than the web browser. MetNet Online provides considerable connectivity for downstream data processing, and it supports several export options, including comma-separated values (CSV), SBML [16] and XGMML. SBML was found sufficient to support all the features contained in the database, and BioPax can be used to encode <annotation>-elements in the output [17, 18]. As it becomes available, we plan to incorporate kinetic data in the SBML files as well. XGMML allows data to be transferred to Cytoscape [19].

Use cases

A horticulturist studies senescence and is interested in ethylene metabolism and signaling [20]. She wants to look at what is known about the process in the model plant, Arabidopsis, and goes to the website and searches for “ethene”, a commonly used synonym. MetNet Online recognizes the synonym and includes “ethylene” in the list of search results. She clicks on the link for the metabolite and sees three Arabidopsis pathways that involve ethylene. She selects all three and creates an integrated view. Although this is helpful in running additional analyses, she does not need the transcription and translation events. She logs into her “My MetNet” account. For easy access, she adds ethylene to her bookmarked entities. Using this shortcut, she visits the three ethylene-related pathways and examines the interactions contained in each

pathway. She bookmarks the interactions that are of particular interest to her. After doing so, she goes back to her bookmark overview page and sees a list of bookmarked interactions, all extracted from the various ethylene-related pathways. She asks for a new integrated view of only these bookmarked interactions. She uses the scaling function in the visualization module to observe the entire network. When she is satisfied, she clicks on the XGMML icon to export her custom network and transfers the data to Cytoscape [19], where she may examine additional properties of the network. The end-result is shown in Figure 5, and the entire scenario is described in more detail in an online video tutorial at http://www.metnetonline.org/tutorial.

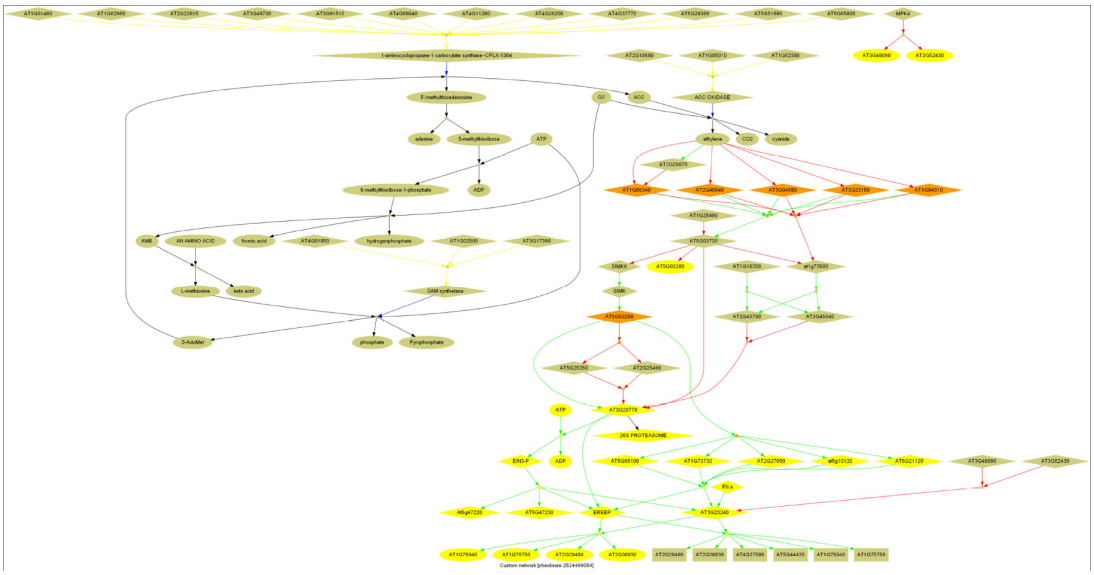

Figure 5: *A custom ethylene-related network. The network was generated dynamically by selecting all pathways in which ethylene (ethane) was found.*

A cell biologist has run a set of microarrays on developing soybean embryos. He identifies a list of differentially (under- and over-) expressed genes. He saves the probe-names as a separate text file (soybean_de.csv) and goes to the MetNet Online website.

He logs into his My MetNet account and creates a personalized list by uploading the text-file. Because the probe-names are recognized by MetNet, he looks for pathways that are over-represented among the differentially expressed genes. Due to the numerous differentially expressed genes (thousands), many pathways show up in an initial quantitative screen. As such, he decides to use the Fisher Exact test module to rank the pathway over-representation by p-value. This presents him with useful information; the list of pathways is still large but they are now ranked by pvalue for relevance. He examines the pathways with the lowest p-values and is thus able to identify other potential gene targets for future experiments and verification. The two use case scenarios for MetNet are described in detail in an online video tutorial at http://www.metnetonline.org/tutorial.

Conclusions

Plant molecular biologists, physiologists, and biotechnologists aim to understand the function of particular genes, polypeptides, or metabolites in plants. Easy and convenient access to integrated information from a variety of biological data repositories is needed to achieve these goals.

We have built a new portal: MetNet Online provides a gateway to integrative systems biology applications. The site generates simple pathways or complex representations of customized interaction sets or combined pathways.

In addition to existing datasets for Arabodopsis, Soybean and Grapevine, MetNet Online incorporates manually curated regulatory events, and also introduces a

subcellular location data layer. Our site supplements other previously created tools. Users can integrate pathways and interactions, and track objects (entities, interactions, and/or pathways) that are of particular interest to them. The site can be reached at http://www.metnetonline.org. Extensive video-tutorials on how to use the site are available through http://www.metnetonline.org/tutorial/.

References

1. Sucaet Y, Deva T. Evolution and applications of plant pathway resources and databases, Briefings in Bioinformatics 2011.

2. Dickerson JA, Berleant D, Cox Z et al. Creating and modeling metabolic and regulatory networks using text mining and fuzzy expert systems. In: Wang J. T. L., Wu C. H., Wang P. eds). Computational Biology and Genome Informatics. Singapore: World Scientific Publishing, 2003, 207-238.

3. Wurtele ES, Li J, Diao L et al. MetNet: Software to Build and Model the Biogenetic Lattice of Arabidopsis, Comp Funct Genomics 2003;4:239-245.

4. Lee EK, Cook D, Wurtele ES et al. (2004), 'GeneGobi : Visual data analysis aid tools for microarray data', COMPSTAT 2004, 16th Symposium of IASC, Physica-Verlag/Springer, Prague.

5. Ding J, Viswanathan K, Berleant D et al. Using the biological taxonomy to access biological literature with PathBinderH, Bioinformatics 2005;21:2560-2562.

6. Yang Y, Engin L, Wurtele ES et al. Integration of metabolic networks and gene expression in virtual reality, Bioinformatics 2005;21:3645-3650.

7. Wurtele ES, Li L, Berleant D et al. MetNet: Systems biology software for Arabidopsis. In: Nikolau B. J., Wurtele E. S. eds). Concepts in plant metabolomics. Springer, 2007, 145-158.

8. Lawrence M, Wickham H, Cook D et al. Extending the GGobi pipeline from R, Computational Statistics 2009;24:195-205.

9. Mentzen WI, Wurtele ES. Regulon organization of Arabidopsis, BMC Plant Biol 2008;8:99.

10. Mentzen WI, Peng J, Ransom N et al. Articulation of three core metabolic processes in Arabidopsis: fatty acid biosynthesis, leucine catabolism and starch metabolism, BMC Plant Biol 2008;8:76.

11. Caspi R, Foerster H, Fulcher CA et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases, Nucleic Acids Res 2008;36:D623-631.

12. Sucaet Y, Wurtele ES. MetNetAPI: A flexible method to access and manipulate biological network data from MetNet, BMC Res Notes 2010;3:312.

13. Arntzen T, Bakken S, Caraveo S et al. PHP: a widely-used general purpose scripting language. http://www.php.net last accessed).

14. Joshi-Tope G, Gillespie M, Vastrik I et al. Reactome: a knowledgebase of biological pathways, Nucleic Acids Res 2005;33:D428-432.

15. Matthews L, Gopinath G, Gillespie M et al. Reactome knowledgebase of human biological pathways and processes, Nucleic Acids Res 2009;37:D619-622.

16. Hucka M, Finney A, Sauro HM et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models, Bioinformatics 2003;19:524-531.

17. Stromback L, Jakoniene V, Tan H et al. Representing, storing and accessing molecular interaction data: a review of models and tools, Brief Bioinform 2006;7:331-338.

18. Stromback L, Hall D, Lambrix P. A review of standards for data exchange within systems biology, Proteomics 2007;7:857-867.

19. Shannon P, Markiel A, Ozier O et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res 2003;13:2498-2504.

20. Meli VS, Ghosh S, Prabha TN et al. Enhancement of fruit shelf life by suppressing N-glycan processing enzymes, Proc Natl Acad Sci U S A 2010;107:2413-2418.

CHAPTER III

METNETAPI

A flexible method to access and manipulate biological network data from MetNet

Convenient programmatic access to different biological databases allows automated integration of scientific knowledge. Many databases support a function to download files or data snapshots, or a webservice that offers “live” data. However, the functionality that a database offers cannot be represented in a static data download file, and webservices may consume considerable computational resources from the host server.

MetNetAPI is a versatile Application Programming Interface (API) to the MetNetDB database. It abstracts, captures and retains operations away from a biological network repository and website. A range of database functions, previously only available online, can be immediately (and independently from the website) applied to a dataset of interest. Data is available in four layers: molecular entities, localized entities (linked to a specific organelle), interactions, and pathways. Navigation between these layers is intuitive (e.g. one can request the molecular entities in a pathway, as well as request in what pathways a specific entity participates). Data retrieval can be customized: Network objects allow the construction of new and integration of existing pathways and interactions, which can be uploaded back to our server. In contrast to webservices, the computational demand on the host server is limited to processing data-related queries only.

An API provides several advantages to a systems biology software platform. MetNetAPI illustrates an interface with a central repository of data that represents the complex interrelationships of a metabolic and regulatory network. As an alternative to data-dumps and webservices, it allows access to a current and "live" database and exposes analytical functions to application developers. Yet it only requires limited resources on the server-side (thin server/fat client setup). The API is available for Java, Microsoft.NET and R programming environments and offers flexible query and broad data- retrieval methods. Data retrieval can be customized to client needs and the API offers a framework to construct and manipulate user-defined networks. The design principles can be used as a template to build programmable interfaces for other biological databases. The API software and tutorials are available at http://www.metnetonline.org/api.

Introduction and background

Analysis of the topology of biological networks provides understanding of structure, function and interaction among cellular entities [1]. As knowledge and understanding of living systems expands, biological network databases are becoming increasingly sophisticated, in terms of data complexity and overall functionality. To facilitate integration with various bioinformatics software packages, many online pathway and network databases offer static data download and conversion methods [2]. Several larger databases, including KEGG [3], BioCyc [4] and Reactome [5], also offer Application Programming Interfaces (APIs). These resources are used by the community

to incrementally enrich datasets, such that each iteration is better and more complete than the previous one.

The MetNet systems biology platform is a suite of software programs that model metabolic and regulatory pathways in plants [6]. At its core is MetNetDB, which represents an integrated pathway-database for plant species and combines various data sources such as AraCyc [7], TAIR [8], AGRIS [9], and atPID [10]. The database allows users to integrate pathways and interactions and keep track of entities of interest in a customizable way.

Through a public website http://www.metnetonline.org, users control various network resources including 1) the ability to bookmark pathways, 2) tracking user-defined components of interest, and 3) a localization datalayer. Users can create new pathways by combining and modifying existing pathways and then save the new pathways. Using the website, pathways can be exported to SBML (for visualization with CellDesigner [11]) or XGMML (for visualization with Cytoscape [12]). All the above functions are now available through our API. MetNetAPI is an alternative interface to the MetNet plaform for tasks that are not easily performed with already existing software tools, time-consuming or repetitive.

An API allows data to be approached and viewed in several modi. Unlike statically-exported files such as data dumps and standardized schema which offer only a single view of the data, whereas an API enables much more user customization, such that a researcher can view or computationally manipulate the data in multiple ways. Consider the static SBML BioModels dataset, wherein each file represents a single

pathway [13]. Assume someone downloads this dataset and wants to gain a more complete understanding of it by creating a list of all the molecular entities that participate in all the pathways. This list can then in turn be used to connect pathways with overlapping components (e.g. pathways in which starch participates can be combined to study starch metabolism and its regulation [14]). However, composing a complete list of entities that make up all the pathways in which starch participates entails writing a piece of custom parsing software. In contrast, an API can implement a method that automatically extracts a list of participating entities for collection of the pathways in which they occur.

Implementation considerations

The choice of an API

Several options exist to share information contained in a biological database. One option for transfer of database content is a data dump. This exposes all the information contained in the database. However, it may require significant effort to understand (and possibly reconstruct) the original database schema.

A second option is to support a standard data format. Chado and BioSQL are two examples of standardized data schema specific to sequence databases [15]; BioPax, SBML and PSI are the most widespread file formats for representation of biological networks [16]. Each standard has its own set of limitations as to what types and resolution of data it can represent. Supporting multiple formats is time-consuming.

A third option is providing an Application Programming Interface (API). A major advantage is that content and functionality are combined [17]. Tying an API directly to a biological database has been done by other groups. MetaCyc [18] is based on the Lisp programming language and interfaces with local MetaCyc-derived databases, while MetNetAPI offers broader programming language support and always connects to the remote "live" MetNetDB database. BioMart [19] is a generic biological repository, and configuring it to support complex network data takes a long time. Its general-purpose nature also makes it slow to run complex queries due to the meta-data that needs to be interpreted first. KEGG [20] and Reactome [21] offer a webservice interface, based on XML and SOAP/REST. A webservice can be considered as a special type of API and provides its own particular problems: A wrapper must be provided around the webservice to facilitate communication and data exchange. This effectively means a secondary API has to be provided to communicate with the initial API. Even as this process can often be automated (through frameworks such as JAX-WS http://jax-ws.dev.java.net/, Axis http://ws.apache.org/ axis/ or XFire http://xfire.codehaus.org/), it is far from efficient. REST-based webservices are somewhat less cumbersome in this regard (they are lightweight, produce human readable results, and require no toolkits like SOAP does), but they have their own peculiarities: Every resource needs to be accessible through a unique URI. This means that information is represented in a hierarchy, which can become complicated very quickly and cumbersome to browse. It is possible, however, to circumvent this problem by allowing querying of the dataset at a different location on the website. The URL to a REST-resource is then a query-string in its own

right. While the messaging protocol involves less overhead than its SOAP-counterpart, the lack of required message meta-data makes these environments at the same time less intuitive and harder to query for complex data. Reactome is one such pathway database [21] that supports REST through BioMart's MartService [19]. Doodle is another resource that supports REST [22], while GenMAPP [23], WikiPathways [24] and CPDB [25] choose to provide SOAP-based services.

If functionality is added to the webservice (either REST or SOAP), supplemental resources - CPU, memory, hard disk - for the server hosting the service must be considered. Webservices therefore seem to be destined to either offer limited functionality (and thus be less useful), or offer extensive functionality but artificially limit access to them because no institution can gather unlimited bandwidth and resources to serve the world. MetNetAPI offers close proximity (strong datatyping) to the MetNetDB database and underlying model, while still able to provide flexibility and abstraction in regard to biological information content. Processing of information mostly occurs on the client running the API, which results in a more distributed load. This presents opportunities to better plan (and distribute) resources across various projects.

API implementation

MetNetAPI is designed as an object model that abstracts and encapsulates the data in the underlying MetNetDB repository. We chose Java, R and Microsoft.NET as target programming languages because they are platform independent and are widely in use today. Users do not need to understand the internal intricacies of the backend database model. The goal is to hide complex data modelling techniques and allow the

bioinformatics software developer (and by extension the biologist) to get started using novel integrated datasets quickly. Overhead is kept to a minimum, as there is no WSDL-file to be parsed, as with SOAP. The structure of the information in MetNetAPI is exposed intuitively through Java reflection mechanisms that are provided in most development environments.

MetNetAPI is a Java jar-file (or .NET Assembly) which contains several logically-ordered namespaces (abstract containers that express semantic categories of code). The main namespace is edu.iastate.metnet (Edu.Iastate.Metnet in .NET). Underlying namespaces and classes allow reasoning by type and allow a programmer to bring biological semantics into the program code. This is in contrast with many other APIs, which result in generic Dictionary-objects, which still require further interpretation and parsing after retrieval. The same argument applies to webservices (especially REST), where the returned output is text-based that requires further processing.

Querying of MetNetDB through MetNetAPI is optimized for efficient memory use. Similar to the Lazy Load concept in the Java persistence library Hibernate http://www.hibernate.org, we adapted Just In Time (JIT) compilation for data retrieval. When retrieving a pathway, only the main data is obtained from the database. Information represented in linked tables (one-to-many or many-to-many) is retrieved when the respective methods are invoked. This occurs transparently, so a client application should function optimally and use a minimal footprint whether retrieving a list of “all pathways”, or constructing an integrated network from “all amino-acid

biosynthesis-related reactions". The JIT data retrieval mechanism not only encapsulates a complex data model, it also makes retrieval and reconstruction of network data efficient. This behaviour is impossible to implement through webservices, as the server cannot "guess" what clients want to do with a returned piece of information in the future. One option would be to provide a verbose-like parameter when calling the webservice, which introduces additional overhead for the programmer consuming the service. Another option would be that the server assumes a worst-case scenario and streams all available (hierarchical) information back to the client, leading to increased (and possibly unnecessary) server-load and network traffic.

Results and deployment

MetNetAPI

MetNetAPI is a flexible API that interacts with and retrieves data from MetNet, an established information resource and suite of software applications for model organisms, currently including Arabidopsis, soybean and grapevine http://www.metnetonline.org [6]. By accessing MetNet infrastructure, the researcher can obtain integrated metabolic and regulatory biological network data, in addition to other new layers of information that were not previously available in any central location.

The API allows a software developer to navigate the database from multiple points of view, without having to understand the underlying database schema. The database can be navigated either as a list of pathways, a list of entities or a collection of organism-centric networks. In contrast, static data files allow only one such point of

view and require customized parsing to determine the answers to specialized questions. Examples of user queries would include “which elements in a list of entities participate in at least two pathways” or “for a given collection of pathways, single out and reconstruct a regulatory network”. MetNetAPI can answer such queries without extensive programming for any respective list of entities (e.g., genes, RNAs, polypeptides, protein complexes, metabolites, or combinations thereof). The API approach allows a database platform to abstract and expose its repository data, along with its functionalities.

Core classes

The MetNetAPI is designed to capture MetNet architecture, which centers around four central classes:

An Entity represents any type of molecular entity that can be found in a biological environment. Entities have a general categorical descriptor that describes the type of an entity, such as “gene”, “RNA” or “Protein Complex”. They can be organism-specific (in the case of a gene) or not (universal metabolites such as ATP or glucose).

A LocalEntity represents a particular entity found within a sub cellular location. An example is the molecule (Entity) ATP, which is found in several compartments (locations) in the cell, including mitochondrion, nucleus, plastid, and cytosol. Therefore, the Entity ATP has four associated LocalEntities.

An Interaction represents the impacts or transformations among entities. Due to the diversity and generalization of the Entity class, Interactions are kept equally generic. Like entities, they are classified. Interactions include enzymatic reactions, transport,

transcription, translation, and various classes of regulatory inhibition and activation such as allosteric effector or indirect positive regulation.

A Pathway represents a group of multiple interactions and the associated biomolecules organized into a convenient functional unit. The pathway concept in MetNetAPI is defined as an unordered collection of Interaction objects. In order to allow developers to determine the start- and end-points of a pathway, getSources() and getSinks() methods are provided.

Peripheral classes are provided to further define pathways and represent MetNet-specific data. The Organism class represents information about organisms currently in MetNetDB. EntityType and InteractionType represent the different types of respective entities and interactions. PathwayClass provides a Pathway Ontology to navigate through the collection of all pathways, which is based on AraCyc pathway classes. CellLocation provides a similar hierarchy that can be used as an alternate pathway ordering tree.

Pathways are arbitrary groupings of interactions. Even for well-defined pathways such as glycolysis and TCA cycle, different views can be created, which may or may not include the genes and the transcriptional and regulatory framework of the various enzymes involved. As more knowledge is acquired through scientific experimentation, pathways may become so complex that it is beneficial to break them into smaller units for some applications. Conversely, smaller pathways may be joined into a larger unit or a super-pathway for meta-analysis.

To model these evolving datasets, a Network class is provided. It serves the purpose of providing custom granularity. A Network object consists of a custom collection of interactions. A Network incorporates the concept of a pathway, yet it is not confined to the boundaries of a predefined pathway. Networks can be constructed either by combining existing pathways or by adding individual interactions.

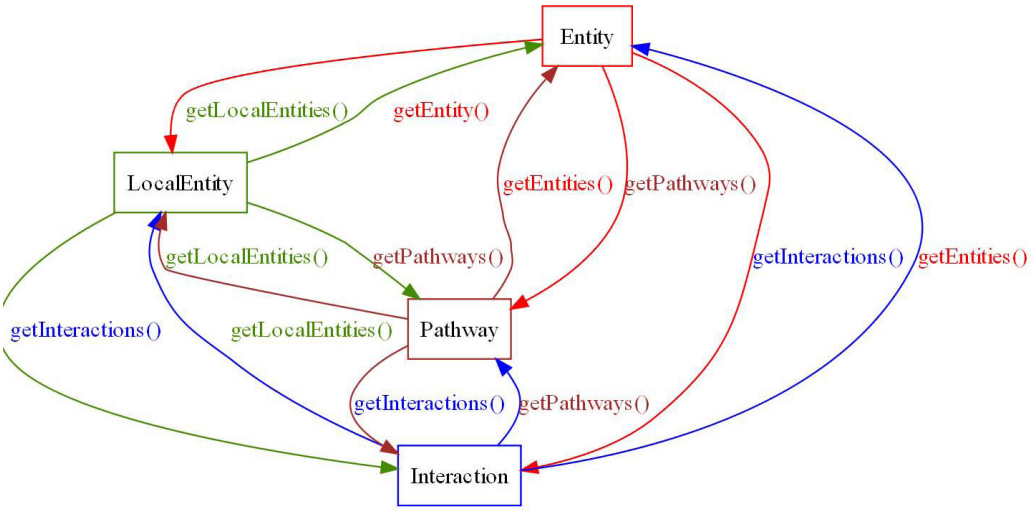

Figure 6: *Interconnectivity between the API's classes. All core classes in MetNetAPI are interconnected. This allows for upward and downward navigation (e.g. one can as easily ask "what entities make up a particular pathway", as "what pathways does a particular entity participate in").*

Several APIs offer top-down approaches to network data. An example is libSBML, in which a pathway consists of reactions, which consist of molecular species [26]. It is currently not possible through libSBML to work backward (e.g. to see which interaction a molecular species participates in). MetNetAPI offers easy navigation and conversion between all its core classes (see Figure 6). This makes it particularly easy to write p-neighbourhood applications, where one is interested in examining the connectedness between network components.

Searching and filtering

Most all network database websites have a search-function. Upon downloading files for offline use, the online functionality is no longer available. This means that a data dump does not always offer the correct amount of information one is interested in. Much effort needs to be invested in study of the original data format and writing parser code to extract the information of interest.

Through MetNetAPI, online search-capabilities are extended and can be integrated in desktop and other applications (these do still need to have networkconnectivity to allow communication between the API and our back-end database). This makes it convenient to execute a large number of queries against MetNet. The investigator can automatically determine which pathways a given list of metabolites participates in, restrict a pathway to its regulatory interactions, or request a list of affected pathways for a set of up-regulated genes. Most Java-classes in the MetNetAPI library have a static search () method, which allows developers to launch queries against MetNetDB in real time, without having to go to a website, fill out a form and submit it.

Filtering using MetNetAPI is similar to searching, but zeros in on results within results. For example, a user could extract all gene regulatory interactions from a previously defined set of pathways (combined as a Network object). Alternatively, a user could look at a complex pathway with 100+ interactions, and decide to remove temporary clutter caused by transcriptional and translational events. The resulting “core”

pathway makes it easier to understand the metabolic functions performed by the pathway.

Applications

The availability of a dynamic code-driven class hierarchy instead of a collection of static, rigid files allows developers to rapidly provide MetNet data and bring its functionality to their own applications. MetNetAPI is object-oriented, which allows for code to be mixed with data (methods and properties). When a collection of pathways is represented by a PathwayVector object, functions to manipulate the member objects are provided. This is preferable to the use of rigid files, or the passing back and forth of Dictionary-like structures.

Source code is provided [Additional file 1] that creates a distance matrix among all 403 pathways in the database. The algorithm results in a GraphViz-compatible http://www.graphviz.org .dot-file, details of which are shown in Figure 7. Additional examples are available on the MetNetAPI tutorial website.

MetNetAPI exports data to standard data formats such as SBML or XGMML (used in Cytoscape). This functionality is available for developers that wish to exploit the richness of MetNetDB. It also allows integration of MetNet-originated data into a more expansive research pipeline. The Network class contains a set of methods that allow export to a variety of standards. To ensure compatibility with a wide spectrum of software, the depth of information has been restricted to a minimum. So, while the Network class is recommended to prepare data for external software such as Jarnac

(SBML) or Cytoscape (XGMML), specialized needs would require a developer to generate customized export-routines.

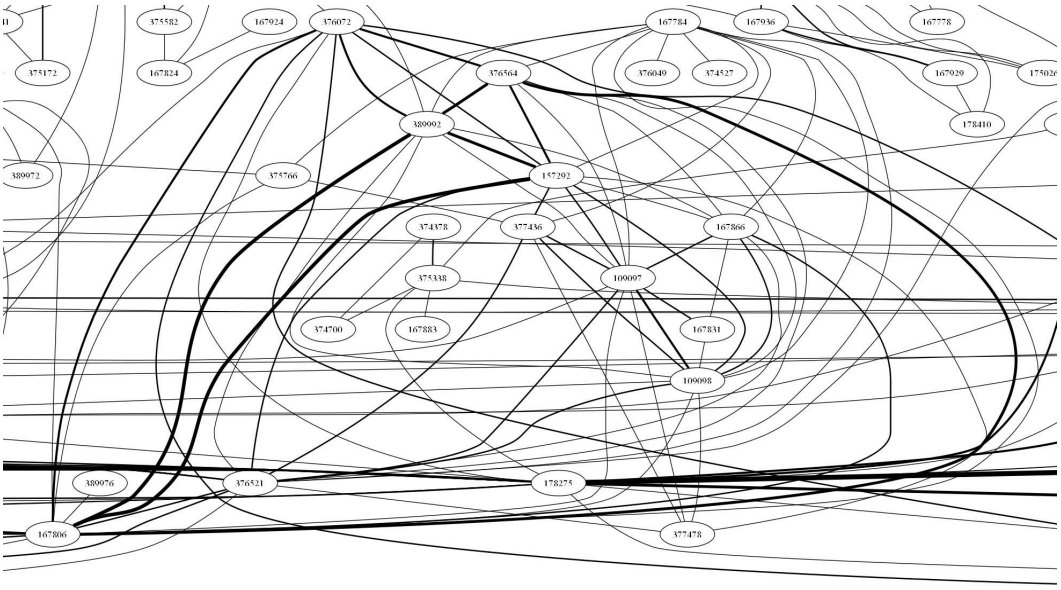

Figure 7: *Details of a map that illustrates shared genes between pathway. With MetNetAPI, it is straightforward to compute a distance matrix between a set of pathways. The matrix can then be visualized with a tool like GraphViz (thicker lines indicate a closer distance). Details of the visualized matrix are shown here.*

MetNetAPI facilitates the creation of static files based on dynamic actions. An example would be to gather the 5 pathways in the database that describe the metabolism and signalling associated with the plant hormones brassinosteroids and auxins into a single Network object, and to export this network to a single XGMML file. This file can be directly imported into Cytoscape to enable visualization and further analysis of a userspecified unit of biology (eliminating the need to import multiple files that represent individual pathways).

Initial adaptations

Several proof-of-concept applications using MetNetAPI have already been developed: We have developed the MetNetScape plugin to allow a user to select an organism and pathway to be imported into Cytoscape. An example of an imported pathway is shown in Figure 8. The plugin is available through our website http://www.metnetonline.org/api/cytoscape/ and source code is available upon request so its functionality may be extended.

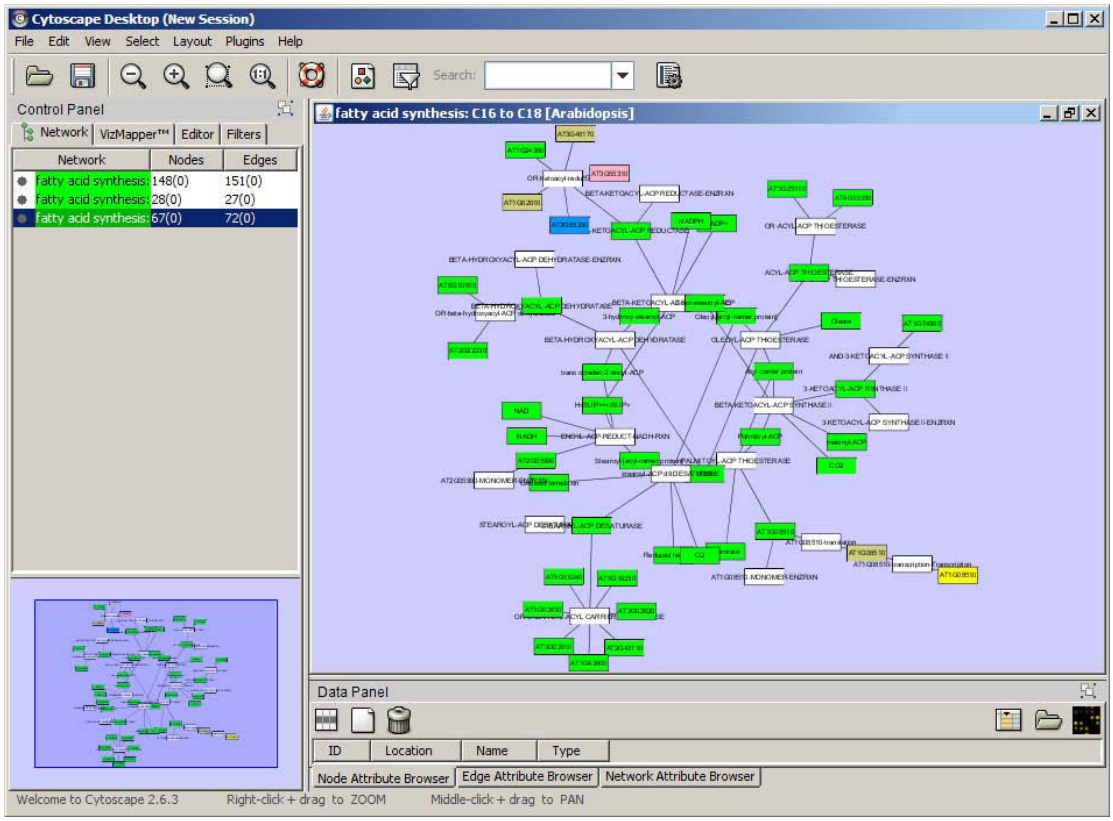

Figure 8: *Cytoscape plugin developed with MetNetAPI. As a proof of concept, a Cytoscape plugin was developed that brings pathway data along with localization information into the Cytoscape environment.*

A more complex plugin has been developed for Cell-Designer [11] to allow exchange and integration of BioCyc and MetNet pathways. The plugin uses the

edu.iastate.metnet.edit namespace to publish new pathways in MetNetDB. This makes MetNet useful as a community annotation platform. The plugin allows for seamless one-click publication of newly constructed pathways into MetNet [27]. It is being used to bring manuallyconstructed grapevine pathways [28] into MetNetDB.

MetaOmGraph (MOG) is an application to display large expression datasets [6,29]. Subsets of entities (genes or metabolites) can be selected in MetNet based on user-specified criteria. These lists can be sent to MOG for further analysis via a user's MetNet profile (a free personal account created through our website [30]). Integration works both ways: genes can be selected in MOG and published to a personal MetNet account [31].

Large biological networks often benefit from visualization in 3D [32]. Walrus http://www.caida.org/tools/visualization/walrus/ is a desktop-application to visualize 3Ddata. A proof-of-concept application has been developed that enables a user visualize MetNet pathways in 3D on a standard computer [33]. The application retrieves data through MetNetAPI to compute the optimal spanning tree to be used by Walrus to create the environment.

MetNetGE [34] is an environment that uses Google Earth infrastructure to produce layered representations of pathways in MetNet. Pathways are visualized as stacked planes, whereby each plane represents a certain type of entity (genes, RNA, polypeptides, or metabolites). MetNetGE uses MetNetAPI to retrieve pathway ontology data and gene information.

Discussion

We have adopted the API as a method to standardize development of applications that exploit the MetNetDB dataset. In addition to facilitating prototyping and rapid application development, this approach ensures consistency across enduser interfaces, command line interfaces, and graphical user interfaces. MetNetAPI is flexible and can be modified, based on needs of internal and external software developers.

We are exploring the possibilities of using the API in environments other than Java. This has already lead to integration of MetNetAPI into Microsoft .NET and R http://www.r-project.org through the rJava bridging software http://www.rforge.net/rJava/.

Advanced programming knowledge (such as SQL or JDBC) is not required for using MetNetAPI. The complexity of the underlying data model is encapsulated within the API. The interface is only slightly less universal than the socket-based protocol provided by BioCyc [18], and the choice of Java allows the API to be used by a broad audience of software developers and bioinformatics researchers. Importantly, unlike a socket-based approach, installation and troubleshooting of MetNetAPI is easy, since it relies on basic Java coding practices. MetNetDB represents a large complex metabolic and regulatory network and contains multiple interaction types, kinetic information, and manually curated subcellular localization assignments.

Conclusions

Online databases often provide data export by means of static downloadable files or dynamic webservices. MetNetAPI provides an additional approach to data export. The API provides a method to standardize development of applications that exploit MetNetDB, but may also serve as a framework and template for other pathway databases. A standardization of terminology among different databases would certainly benefit developers that work on integrative applications. Many databases expose similar types of data, and the definition of a minimal set of interfaces that pathway database APIs may be expected to implement would be helpful. MetNetAPI can be a first step in this direction.

Apart from facilitating prototyping and rapid application development, our approach ensures consistency and data integrity across command line interfaces and graphical user interfaces alike. The choice of Java and Microsoft. NET allows the API to be used by a broad audience of software developers and bioinformaticists. The complexity of the underlying data model is encapsulated within the API. Because it is a Java-API rather than a webservice, more functionality can be provided without requiring extensive computational resources on the server-side.

For a densely populated and information-rich database (such as MetNetDB), our API model offers many advantages. It has the ability to incorporate online search capabilities into custom-built applications. It also offers the option to customize the granularity of pathways of interest.

MetNetAPI captures user-defined network structures into self-contained semantic objects. Through Network objects, combinations of existing or putative novel pathways can easily be constructed, manipulated and refined. MetNet is an information resource, as well as an active toolkit to develop new hypotheses. Many complicated operations, which would be difficult to implement via xml or text-based files, can be accomplished through MetNetAPI. These feature flexible capabilities to agglomerate data over multiple pathways, to examine connectivity among different datatypes, and prepare custom datasets for use in other downstream applications. MetNetAPI is fully documented, free of charge and can be downloaded from http://www.metnetonline.org/api/cytoscape/.

References

1. Steuer R, Lopez GZ: Global network properties. In Analysis of biological networks. Edited by: Junker BH, Schreiber F. Hoboken, NJ: John Wiley 2008:31-64, [Pan Y, Zomaya AY (Series Editor): Bioinformatics: Computational techniques and engineering].

2. Suderman M, Hallett M: Tools for visually exploring biological networks. Bioinformatics (Oxford, England) 2007, 23:2651-2659.

3. Kawashima S, Katayama T, Sato Y, Kanehisa M: KEGG API: A Web Service Using SOAP/WSDL to Access the KEGG System. Genome Informatics 2003, 14:673-674.

4. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, et al: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic acids research 2010, 38:D473-479.

5. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, et al: Reactome: a knowledge base of biologic pathways and processes. Genome Biol 2007, 8:R39.

6. Wurtele ES, Li L, Berleant D, Cook D, Dickerson JA, Ding J, Hofmann H, Lawrence M, Lee EK, Li J, et al: MetNet: Systems biology software for Arabidopsis. In Concepts in plant metabolomics. Edited by: Nikolau BJ. Wurtele ES: Springer; 2007:145-158.

7. Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY: MetaCyc and AraCyc. Metabolic pathway databases for plant research. Plant physiology 2005, 138:27-37.

8. Poole RL: The TAIR database. Methods in molecular biology 2007, 406:179-212.

9. Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E: AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC bioinformatics 2003, 4:25.

10. Cui J, Li P, Li G, Xu F, Zhao C, Li Y, Yang Z, Wang G, Yu Q, Li Y, Shi T: AtPID: Arabidopsis thaliana protein interactome database–an integrative platform for plant systems biology. Nucleic acids research 2008, 36: D999-1008.

11. Funahashi A, Tanimura N, Morohashi M, Kitano H: CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. BIOSILICO 2003, 1:159-162.

12. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003, 13:2498-2504.

13. Le Novere N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, et al: BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic acids research 2006, 34: D689-691.

14. Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES: Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. Plant J 2009, 58:485-498.

15. Stein LD: Integrating biological databases. Nature reviews 2003, 4:337-345.

16. Stromback L, Lambrix P: Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. Bioinformatics (Oxford, England) 2005, 21:4401-4407.

17. Cohen Y, Feldman YA: Automatic high-quality reengineering of database programs by abstraction, transformation and reimplementation. ACM Transactions on Software Engineering and Methodology 2003, 12:285-316.

18. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, et al: The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic acids research 2008, 36:D623-631.

19. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: BioMart–biological queries made easy. BMC Genomics 2009, 10:22.

20. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: The KEGG resource for deciphering the genome. Nucleic acids research 2004, 32:D277-280.

21. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, et al: Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 2009, 37:D619-622.

22. Marino-Ramirez L, Minor JL, Reading N, Hu JC: Identification and mapping of self-assembling protein domains encoded by the Escherichia coli K-12 genome by use of lambda repressor fusions. J Bacteriol 2004, 186:1311-1319.

23. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, Dahlquist KD, Doniger SW, Stuart J, Conklin BR, Pico AR: GenMAPP 2: new features and resources for pathway analysis. BMC bioinformatics 2007, 8:217.

24. Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, Conklin BR: Mining biological pathways using WikiPathways web services. PloS one 2009, 4:e6447.

25. Kamburov A, Wierling C, Lehrach H, Herwig R: ConsensusPathDB–a database for integrating human functional interaction networks. Nucleic Acids Res 2009, 37:D623-628.

26. Bornstein BJ, Keating SM, Jouraku A, Hucka M: LibSBML: an API library for SBML. Bioinformatics (Oxford, England) 2008, 24:880-881.

27. Pathway data integration between datasources via CellDesigner. [http://www.public.iastate.edu/~jlv/celldesignerplugins.shtml].

28. Grimplet J, Cramer GR, Dickerson JA, Mathiason K, Van Hemert J, Fennell AY: VitisNet: “Omics” Integration through Grapevine Molecular Networks. PloS one 2009, 4:e8365.

29. Mentzen WI, Wurtele ES: Regulon organization of Arabidopsis. BMC plant biology 2008, 8:99.

30. MetNet Online. [http://www.metnetonline.org].

31. MetaOmGraph. [http://www.metnetdb.org/MetNet_MetaOmGraph.htm].

32. Yang Y, Engin L, Wurtele ES, Cruz-Neira C, Dickerson JA: Integration of metabolic networks and gene expression in virtual reality. Bioinformatics (Oxford, England) 2005, 21:3645-3650.

33. Tools and ideas for visualizing Systems Biology data in 3D. [http://vrac.iastate.edu/~jlv/3D/].

34. Ming J, Swaminathan S, Wurtele ES, Dickerson JA: MetNetGE: Visualizing biological networks in hierarchical views and 3D tiered layouts. IEEE International Conference on Bioinformatics and Biomedicine Workshop; Washington, DC 2009, 287-294.

CHAPTER IV

CANSTOREX

A framework for collaboration and remote access into XML

The proliferation of XML-formatted data presents challenges for data storage methodologies: size, heterogeneity and sparseness are difficult problems to solve with existing database platforms. CanStoreX is a new in-house built XML storage solution that addresses these issues. Using novel XML pagination technology, large volumes of data are efficiently managed. Technologies such as DOM (Document Object Model) and XQuery (XML Query) enable the practical use of XML. Both DOM and XQuery are implemented on top of the CanStoreX storage engine. Middleware was developed for connectivity and to allow integration of the solution into a multi-tier application development paradigm. Finally, CanStoreX is applied to a real-life biological dataset of heterogeneous disparate data. The case study illustrates how pure XML solutions can significantly simplify problems that are complex to solve with conventional techniques. The application shows that CanStoreX is a stable solution to consider for building novel XML repositories.

Introduction and background

The Extensible Markup Language – XML – is a flexible format for storage, access and exchange of semi-structured data in a variety of applications (Bray et al., 2006). Since inception in 1998, XML has gained broad adoption. The technology can be

described as a basic text-file format, on top of which domain-specific data-formats can be built. Implementations can be found in diverse areas such as manufacturing (PSI/XML; Lubell and Schlenoff, 1999), chemistry (CML; Kuhn et al., 2007), healthcare (Daniel-Le Bozec et al., 2006) and systems biology (SBML and Biopax; Strömbäck et al., 2006). All these and other XML sub-formats are described as XML-documents in a meta-language of Document Type Definitions – DTD – or XML Schema.

An efficient approach is needed to store, retrieve, query and update large XML datasets. Various approaches have been developed to interact with XML data. The two most widely used methods for access are the Simple API for XML – SAX – and the Document Object Model – DOM. SAX is an event-driven API that scans a document from start to finish. Upon finding matched events, predefined actions are taken (Megginson, 2001). In contrast, DOM parses and maps an XML document to an internal tree structure reflecting the hierarchical structure of the document (Le Hors et al., 2004). While the SAX parser is efficient, it is difficult to use it to exploit the hierarchical tree structure of XML documents in the way the DOM parser does. DOM however requires the whole document to be loaded into memory as a fully expressed tree. Memory limitations therefore have become a major issue in DOM applications and are limited to documents that are at most 10-20 Megabytes in size in current systems. To illustrate: opening a 12 Megabyte XML-file in Internet Explorer 8 beta 2 consumes 1 Gigabyte of memory. In addition to parsing technologies, querying of XML documents is also possible. The most recent standard method in this respect is XQuery (Boag et al., 2007).

The integration of XML-data in legacy and new applications can occur at different levels. One possibility is to use XML solely as a communication protocol between hosts. Each host functions independently, e.g. by parsing each message into a relational database (Florescu and Kossmann, 1999) or by storing each incoming/outgoing message unprocessed as a plain text file. Depending on the requirements, this basic approach may be sufficient, yet recently more advanced solutions have become available, both commercially and experimentally.

In SQLServer 2000, Microsoft extended its own T-SQL format with a new FOR XML clause to allow relational data to be exported and formatted as XML (Rys, 2001). Interaction with XML data is also available through other vendors such as IBM (XML Extender for DB2 ; Cheng and Chu, 2000) or Oracle (Murthy and Banerjee, 2003). The products are typically referred to as XML Enabled databases (XenDB).

One step further up are “pure” or Native XML databases (NXD). Such databases don’t map XML-data internally to conventional paradigms such as RDBMS. Several NXDs are currently available, including Tamino (Schöning and Wäsch, 2000), X-Hive (http://www.x-hive.com), Xyleme (Aguilera et al., 2000), Natix (Fiebig et al., 2002), TIMBER (Jagadish et al., 2002), Berkeley DB XML (Sleepycat software, 2003), and eXist (Meier, 2003). Natix is a subtree-based strategy. It divides the XML document tree into subtrees according to the physical page size, so that each subtree is a record. A split matrix is defined to ensure that correlated elements remain clustered. TIMBER is a native XML database system built on Shore (Carey et al., 1994) storage manager. Both

systems utilize the element-based storage strategy, where each element is an atomic unit in the storage and is organized in a pre-ordered manner.

Managing XML data poses several unique challenges. First, when choosing XML as a logical storage format, the files can become very large. For example, OpenStreetMap.org makes its data available as a 4.6 Gigabyte compressed file; uncompressed data is 95 Gigabytes (Haklay and Weber, 2008). In OpenStreetMap's case, handling the large volume of data is overcome by offering scripts to convert the data to an RDBMS structure.

Another issue is the heterogeneity of different XML-files. While conversion to a relational format is certainly possible for highly structured data, this is often a sub-optimal solution for datasets that contain sparse and/or heterogeneous data. In addition, domain-specific information may be stored over a set of XML files that need not all adhere to the same XML DTD or schema.

A number of scenarios are given by using SBML (Hucka et al., 2003) and BioPax (Bader et al., 2006) as examples; two formats that are both used to store biological pathway information:

A researcher may be interested in two pathways, A and B. Yet A is only available in SBML format, whereas B is in BioPax. Conversion of one format into another would lead to loss of certain format-specific data, and storage in two different format-specific databases complicates querying.

A researcher has an existing collection of pathways, and wants to update his repository. He discovers that the updated files are in SBML Level 2 format, where his

current relational model incorporates the SBML Level 1 format. Several fundamental changes have been made in Level 2, which require significant remodeling of the relational model.

A researcher downloads a third-party set of pathways from a website, yet finds that the information contained in the files is vastly diverse: some pathways contain information down to the molecular level, other pathways contain literature references. While all files are legitimate SBML-files, modeling the contained diversity results in a sparse relational database with complex joins.

BioMart is an example of how different data-formats from different sources can be integrated. For each file, an import-filter converts the original format to a custom relational database structure. This back-end database then contains all data and allows querying through an integrated interface (Durinck et al., 2005). Commercial applications exist as well, and Güler et al. (2003) use Microsoft Biztalk to achieve this. However all these proposals are work-arounds to the general problem of data integration from heterogeneous sources. We present a new NXD solution: Canonical Storage of XML (CanStoreX). The remainder of this manuscript discusses implementation issues with the platform as well as how CanStoreX overcomes the aforementioned issues regarding management and storage of XML-data. Application of CanStoreX is illustrated by means of a heterogeneous dataset of biological pathways obtained from the European Molecular Biology Laboratory (Le Novère et al., 2006).

Materials and methods

CanStoreX

CanStoreX is a novel in-house native XML storage and database management system. It uses a tree-based storage strategy, and stores an XML document according to its original hierarchical structure. CanStoreX breaks an XML document into pages. Each page is a self contained XML document and is linked with other pages through inter-page references. Thus, to process an element in the XML document, the system only needs to load one page into memory at a time. A new DOM API (CanStoreX DOM) was built to support tree-like processing of stored XML documents. An XQuery engine has been implemented on top of CanStoreX DOM.

Charon connectivity layer

While an isolated database can certainly be useful for individual use, in practice one typically wants a more scalable model whereby a central datastore can be consulted by multiple users locally and remotely. Additionally, a server architecture allows for different clients to obtain data in a preferential manner (e.g. web-application, Java applet, and .Net desktop application) and further manipulate and format resulting data to satisfy desired output requirements.

As it sits effectively between client and server, Charon can be considered middleware. It presents a connectivity layer that effectively transforms CanStoreX into a TCP/IP server. Charon thus facilitates the creation of drivers and other client applications, similar to JDBC, ODBC or OLEDB (Abdullat, 2004). These architectural data exchange frameworks offer client applications a common interface to datastores,

regardless of underlying (database) server specifics. While not currently implemented, Charon allows a driver to be developed and integrated into a JDBC, ODBC or OLEDB architecture.

Design goals for Charon are twofold: it has to be well-defined (keywords and protocol) and client-independent (operating system and programming language). The interface goes beyond simple linear streaming typical in relational databases and allows clients to navigate an XML document using CanStoreX DOM while only materializing the portions that are needed.

BioModels dataset

The BioModels dataset consists of a set of curated and non-curated biological pathways. It is available for download through the European Molecular Biology Laboratory (http://www.embl.org; Le Novère et al., 2006). Each pathway is contained in an individual SBML (Hucka et al., 2003) file. In order to facilitate integrated research, it is desirable to have a mechanism in place that can integrate queries over all files simultaneously. However, a simple collated XML-file is too complex to handle with traditional XML DOM. Therefore, the integrated XML-file was uploaded to CanStoreX. The CanStoreX XQuery implementation was then used to query the entire dataset as a single entity, rather than looping over a set of individual files.

Results and implementation

The overall architecture of CanStoreX consists of five key components: Disk Space Manager, Buffer Manager, Loading Engine, CanStoreX DOM API and XQuery Engine.

The Disk Space Manager manages the local storage for CanStoreX. It supports the concept that a page is a unit of data and provides commands to allocate, deallocate, read or write pages. The size of a buffer residing in main memory is chosen to be the same as the size of a page on the disk, such that a reading or writing operation can be completed in one disk I/O. The Buffer Manager manages a pool of buffers. It is responsible for bringing pages from disk to main memory and back as needed.

The Load Engine parses and loads an original XML document into the CanStoreX storage using a pagination algorithm (Ma et al.,2004; Patenroi,2005). It parses an XML document and adds storage-facilitating nodes on the fly to support pagination. Pages are interconnected to reflect the structure of the document. The end result of pagination would be a group of pages in the storage, rooted at the root page of the XML document. The document is permanently stored in a ready to use form.

The CanStoreX DOM API allows users to navigate within the DOM tree, which corresponds to an XML document in the storage. With CanStoreX DOM API, the parts of a document needed by the user are automatically brought into main memory. The XQuery Engine processes XQuery expressions (Boag et al., 2007) from users, and communicates with the CanStoreX DOM API to access the DOM tree of an XML document.

Individual SBML-files were concatenated into a single XML-file and imported into CanStoreX. The XQuery Engine was then used to examine the integrated dataset. The total number of molecular species over all models is 2,139. The number of unique molecular species is 1,640. Among the most observed molecules are ATP and NADH, which are used in energy transfer in the cell. Other molecules that are present over multiple pathways are p53, a protein involved and targeted in cancer research (Staples et al., 2008). MAPK and MEK are two kinases that are well-documented as signaling molecules (e.g. Ballif and Blenis, 2001). Molecular species occurring in two or more pathways are the more interesting targets of scrutiny, because they allow for the coupling and integration of different pathways.

Another application of finding similar elements across different pathways is the integration of their respective local annotations. All pathways are stored in SBML, wherein each molecular species can have a generic <annotation> element. The element can be populated by any legal XML-code. This results in both sparse and heterogeneous datasets, at least when considering the data in a conventional relational paradigm. To resolve ambiguity in the meaning of the information, <annotation>-elements are associated with namespace declarations. These depend on the host application that generated the pathway (SBML is supported by around 100 different applications). Different applications can therefore store different information for molecules. For example, one application may store binding sites for a protein P in pathway A while a different application may store the amino acid sequence for protein P in pathway B. Integration is now beneficial because we can acquire both the binding sites and the

amino acid sequence for protein P (as well as possibly integrate pathways A and B, cf. supra).

Charon offers the software designer many options to utilize CanStoreX in an application. Through a socket communication protocol, a programmer has the option to receive a stream of XML data, or use a hierarchical DOM-like approach of the data. The eventual choice depends on the type of application (Desktop vs. Web) and the size of the expected return stream (Megabytes vs. Gigabytes). Charon is a fundamental component of CanStoreX in order to allow implementation into scalable multi-tier applications.

Discussion

CanStoreX is a novel scalable Native XML Database – NXD. In separate tests, the database has shown to be able to handle up to 100 Gigabyte test-documents generated by the XMark benchmarking tool (Schmidt et al., 2001). This manuscript describes results of applying CanStoreX to a real-life dataset of biological pathways. The technology enables the integration of data collected from multiple sources into a single seamless XML document and data repository.

CanStoreX offers obvious physical advantages, such as reduced memory requirements and scalability. In addition, one of the biggest advantages offered is the possibility to query all data simultaneously. While it is technically possible to loop over a set of smaller SBML/XML-files individually, applying a single query to the entire dataset is much more convenient and faster (both in designing the query and its execution).

Even loading the dataset in a regular web-browser / viewer required 1 GigaByte in memory. Specialized software (WMHelp XMLPad Pro; http://www.wmhelp.com) still consumed 150 MegaBytes, a more than ten-fold increase compared to the actual filesize. These technologies only allow viewing the data and do not support any query capability. Therefore, the use of an XML database solution is justified.

We have already shown in the result section set that the SBML format by design results in both sparse and heterogeneous datasets. XML is the best solution to handle such data, since only the XML elements that are present are stored (empty blocks do not translate to relational NULL values). Different versions of the format are no more problematic either, as CanStoreX does not require a predefined metadata structure to map data to. All that it expects is well-formatted XML.

Integration of biological pathways in CanStoreX is more than a technical exercise. Having all pathways available as a single queryable unit leads to new applications. Questions can now be asked that were very difficult to ask with conventional relational or DOM technology. Examples of this are the coupling of pathways based on common molecular species membership, and the composition of integrated annotation for selected species.

A problem that the current integrated dataset suffers from is knowledge and recognition of synonyms. E.g. “Fructose 1,6-bisphosphate” and “Fructose 1,6-phosphate” identify the same chemical component, yet today they show up as two individual entities, each occurring twice. The problem is systematic, since chemical compounds inherently have a number of designations, such as CAS-number, systematic

name, and structural formula. A similar case is a reference to “water” and “H2O”. The problem runs deeper than simple curation, since an end-user still expects to be able to search for both terms. The result should be all instances of pathways that include either “H2O” or “water”. Solving this problem will allow the construction of additional integrated super-pathways. Once synonyms are resolved, it would certainly be possible to standardize naming in search results.

The current XQuery engine in CanStoreX only supports a subset of the grammar that includes the basic FLWR expressions, path expressions and a few basic operators and functions. The updates operations are expected to be supported by the XQuery engine in CanStoreX in the near future.

Finally, this case study looks only at a limited set of pathways available in EMBL’s BioModels repository. Only 21% (337) of all molecular species are present in two or more pathways (although this number can be expected to somewhat increase when solving the aforementioned synonym problem). Plans for the future therefore include obtaining and integrating pathways from additional sources such as KEGG (Kanehisa, 2002) or Reactome (Joshi-Tope et al., 2005).

Conclusion

As a general data format, XML has been very successful. The omni-presence of XML datasets today leads to new challenges in storage and querying technology. CanStoreX is a novel platform that is both scalable and flexible. Aside from benchmarking CanStoreX internally with industry-accepted methodologies (XMark;

Schmidt et al., 2001), we also examined a real-life biological hierarchical heterogeneous dataset. The dataset is hard to manipulate or query without the help of CanStoreX, and the platform is at the stage where it can be easily integrated into a variety of software development paradigms thanks to the Charon communication layer. After loading the BioModels dataset, we also find that the availability of integrated pathway data leads to new questions that are hard to answer without CanStoreX. Feedback from biologists in months to come will undoubtedly result in additional functionality.

XML is very flexible at many levels and having in-house technology helps in solving problems on supporting collaboration and remote access. Unlike relational databases that are a representation of the data in tabular form, XML offers more options to support new applications. Specifically, we see tremendous opportunity for future expansion in the area of heterogeneous and sparse data, as illustrated here by the BioModels dataset. Further potential for CanStoreX exists in the field of collaboration (e.g. introduction of new problem-specific tags to facilitate versioning), an important area in a world with ever expanding project teams.

References

[1] Abdullat A, 2004. Internet and web-based database technology. Information Systems Education Journal 2(17): 1-10.

[2] Aguilera V, Cluet S, Veltri P, Vodislav D, Wattez F, 2000. Querying XML Documents in Xyleme. In Proceedings of the ACM-SIGIR Workshop on XML and Information Retrieval.

[3] Bader GD, Cary M, Sander C, 2006. BioPAX – Biological Pathway Data Exchange Format. Encyclopedia of Genomics, Proteomics and Bioinformatics, New York: John Wiley & Sons, Ltd.

[4] Boag S, Chamberlin D, Fernandez MF, Florescu D, Robie J, Simeon J, 2007. XQuery 1.0: An XML query language. Technical report, World Wide Web Consortium, 2007. W3C recommendation 23 January 2007.

[5] Bray T, Paoli J, Sperberg-McQueen CM, Maler E, Yergeau F, 2006. Extensible Markup Language (XML) 1.0 (Fourth Edition) W3C recommendation, 16 August 2006.

[6] Carey MJ, DeWitt DJ, Franklin MJ, Hall NE, McAuliffe ML, Naughton JF, Schuh DT, Solomon MH, Tan CK, Tsatalos OG, White SJ, Zwilling MJ, 1994. Shoring up Persistent Applications. In Proc SIGMOD Conference: 383-394.

[7] Cheng J, Xu J, 2000. XML and DB2. 16th International Conference on Data Engineering (ICDE'00), 569

[8] Daniel-Le Bozec C, Henin D, Fabiani B, Bourquard K, Ouagne D, Degoulet P, Jaulent MC, 2006. Integrating anatomical pathology to the healthcare enterprise. Studies in health technology and informatics 124: 371-376.

[9] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W, 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21(16): 3439-3440.

[10] Fiebig T, Helmer S, Kanne C-C, Mildenberger J, Moerkotte G, Schiele R, Westmann R, 2002. Anatomy of a Native XML Base Management System. Technical Report TR-02-001, Universität Mannheim.

[11] Florescu D, Kossmann D, 1999. Storing and querying XML data using an RDMBS. IEEE Data Eng. Bull 22(3): 27-34.

[12] Güler S, Eberhart A, Rojas I, 2003. Web-based exchange of biochemical information. Bioinformatics 19(13), 1730-1731.

[13] Haklay M, Weber P, 2008. OpenStreetMap: User-Generated Street Maps. Pervasive Computing, IEEE 7(4): 12-18.

[14] Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J, 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19, 524–531.

[15] Jagadish HV, Al-Khalifa S, Chapman A, Lakshmanan LVS, Nierman A, Paparizos S, Patel JM, Srivastava D, Wiwatwattana N, Wu Y, Yu C, 2002. TIMBER: A native XML database, The VLDB Journal 11(4): 274-291.

[16] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L, 2005. Reactome: a knowledgebase of biological pathways. Nucleic Acids Research 33(Database issue):D428-432.

[17] Kanehisa M, 2002. The KEGG database. Novartis Found Symp. 247:91-101; discussion 101-3, 119-28, 244-52.

[18] Kuhn S, Helmus T, Lancashire RJ, Murray-Rust P, Rzepa HS, Steinbeck C, Willighagen EL, 2007. Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML vocabulary for spectral data. Journal of chemical information and modeling 47(6): 2015-2034.

[19] Le Hors A, Le Hegaret P, Wood L, Nicol G, Robie J, Champion M, Byrne S, 2004. Document Object Model (DOM) level 3 core specification. Technical report, World Wide Web consortium (W3C) recommendation 07 April 2004.

[20] Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M, 2006. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Research 34(Database issue):D689-691.

[21] Lubell J, Schlenoff C, 1999. Process Representation Using Architectural Forms: Accentuating the Positive, Markup Technologies ’99.

[22] Ma S, Gadia SK, Berleant D, Huang X, 2004. Implementation of a canonical native storage for XML. Master’s Thesis. Department of Computer Science. Iowa State University, 2004.

[23] Megginson D, 2001. SAX: A Simple API for XML. Technical Report, Megginson Technologies, http://www.saxproject.org/.

[24] Meier W, 2003. eXist: An Open Source Native XML Database. In Web, Web-Services, and Database Systems, Springer, Berlin / Heidelberg: 169-183.

[25] Murthy R, Banerjee S, 2003. Xml schemas in Oracle XML DB. Proceedings of the 29th international conference on very large databases 29: 1009-1018.

[26] Patanroi D, 2005. Binary page implementation of a canonical native storage for XML. Master’s Thesis. Department of Computer Science. Iowa State University, 2005.

[27] Rys, M, 2001. Bringing the Internet to your database: using SQL server 2000 and XML to build loosely-coupled systems. Data Engineering, 2001. Proceedings. 17th International Conference on 26 April 2001: 465 – 472.

[28] Schmidt AR, Waas F, Kersten ML, Florescu IM, Carey MJ, Busse R, 2001. The xml benchmark project. Tech. rep., Technical Report INS-R0103, CWI, Amsterdam, The Netherlands, April.

[29] Schöning H, Wäsch J, 2000. Tamino - An Internet Database System. In C. Zaniolo, P. C. Lockemann, M. H. Scholl, and T. Grust, editors, Advances in Database Technology – EDBT 2000, Proc. 6th Int. Conf. on Extending Database Technology, volume 1777 of Lecture Notes in Computer Science, Springer: 383–387.

[30] Sleepycat software, 2003. Berkely DB XML. http://www.sleepycat.com/products/xml.shtml

[31] Staples OD, Steele RJ, Lain S, 2008. p53 as a therapeutic target. Surgeon. 6(4): 240-243.

[32]Strömbäck L, Jakoniene V, Tan H, Lambrix P, 2006. Representing, storing and accessing molecular interaction data: a review of models and tools. Briefings in bioinformatics 7(4): 331-338.