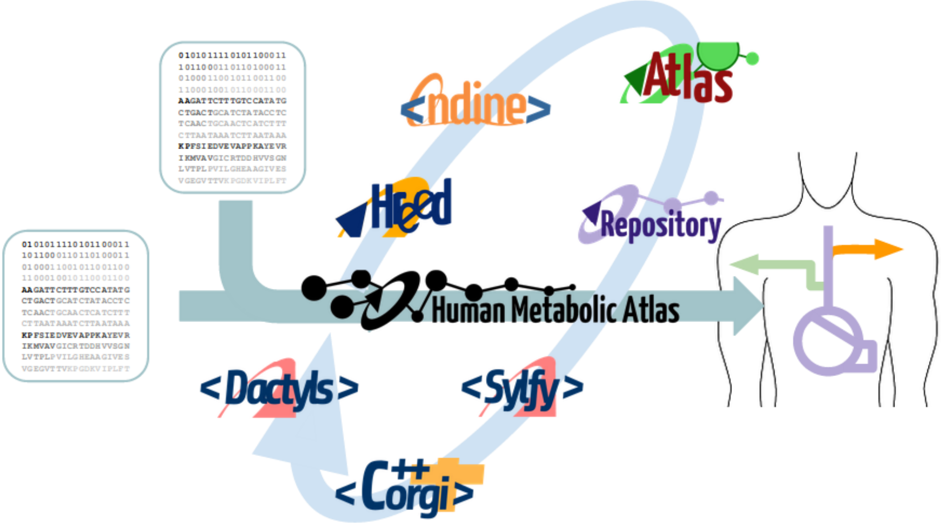
CHALMERS



Database and Visualization for Advanced Systems Biology

NATAPOL PORNPUTTAPONG

Department of Chemical and Biological Engineering
Systems and Synthetic Biology
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2014

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN

Database and Visualization for Advanced Systems Biology

NATAPOL PORNPUTTAPONG

Department of Chemical and Biological Engineering
Systems and Synthetic Biology
CHALMERS UNIVERSITY OF TECHNOLOGY

Göteborg, Sweden 2014

Database and Visualization for Advanced Systems Biology

NATAPOL PORNPUTTAPONG
ISBN 978-91-7385-983-7

© NATAPOL PORNPUTTAPONG, 2014

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr. 3664
ISSN 0346-718X
Department of Chemical and Biological Engineering
Systems and Synthetic Biology
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: +46 (0)31-772 1000

Cover:
Front
Data utilization process prior to human metabolism
by Natapol Pornputtapong
Back
by Kwanjeera Wanichthanarak

Chalmers Reproservice
Göteborg, Sweden 2014

Database and Visualization for Advanced Systems Biology

Thesis for the degree of Doctor of Philosophy in
NATAPOL PORNPUTTAPONG
Department of Chemical and Biological Engineering
Systems and Synthetic Biology
Chalmers University of Technology

Abstract

In the information age, there is plenty of information available publicly in the field of biology. Utilization of biological data is still slow and inefficient compared to the amount of data generated. This problem arise due to the specific characteristics of biological data, which are complex, dynamic and variable. With the introduction of high throughput technologies, the gap between data creation and utilization has become wider. This issue is critical and poses a challenge in the field of systems biology, where data from several sources are needed for model construction and analysis.

In order to build a data ecosystem to support human tissue specific genome reconstruction and further analysis, a collection of libraries, applications and a web site have been developed. A dedicated database management system was designed specifically for metabolic and related data to support human tissue specific genome scale metabolic model reconstruction providing data standardization and data integration. Two database APIs, Corgi and Dactyls, were developed following the Object-oriented data model to fulfill the database management system's functions. This database management system was used to manage, provide and exchange information concerning particularly human metabolism. Furthermore was developed the visualization system, Ondine that allows overlaying of data and information on metabolic pathway maps with a zoom/pan user interface.

In order to efficiently deploy human tissue specific metabolic information from a collection of genome-scale metabolic models (GEMs), the Human Metabolic Atlas (HMA) website was created as an online resource to provide comprehensive human metabolic information as models and as a database for further specific analysis. In addition, the Atlas also serves as a tool for communicating with the wider research community. The Atlas, providing a visualization of the metabolic map implemented on the Ondine engine, provides comparative information of metabolism among deposited GEMs. Hreed is intended to provide accurate information about human metabolism in order to exchange data with the community and to support metabolic network based modeling and analysis through both the graphical and application programming interfaces. This data ecosystem development and implementation is the starting step for the enhancement of data utilization in systems biology.

Keywords: database design; database system; omic data visualization system; data integration; data standardization

List of publications

This thesis is based on the work contained in the following publications.

I Pornputtapong, N., J. Nielsen, and I. Nookaew (2014). Ondine: A web application for multilevel omics data integration and visualization, submitted

II Pornputtapong, N., K. Wanichthanarak, A. Nilsson, I. Nookaew, and J. Nielsen (2014). A dedicated database system for handling multi-level data in systems biology, submitted

III Agren, R., S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, J. Nielsen (2012). Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT ed. C.D. Maranas. PLoS Comput Biol 8: e1002518. http://dx.plos.org/10.1371/journal.pcbi.1002518.

IV Pornputtapong, N., I. Nookaew, and J. Nielsen (2014). Human Metabolic Atlas: a web resource for human metabolism

Additional publication not included in this thesis:

V Nookaew, I., M. Papini, N. Pornputtapong, G. Scalcinati, L. Fagerberg, M. Uhlèn, J. Nielsen (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res 40: 10084–97. http://nar.oxfordjournals.org/content/40/20/10084 (Accessed January 12, 2014).

Contributions

I Designed and developed the web application and visualization engine. Drafted and edited the paper.

II Designed and developed the database system. Participated in test case design. Drafted and edited the paper.

III Designed and developed the database system. Performed data propagation. Drafted and edited the paper.

IV Designed the web site and developed the database API library. Performed data propagation. Drafted and edited the paper.

Additional publication not included in this thesis:

V Performed *de novo* assembly of RNA sequencing data and sequence variation finding. Implemented the genome viewer.

Contents

List of figures

List of tables

To my dear parents and siblings.
To my girlfriend, Charuwan.

”It is possible to commit no errors and still lose. That is not a weakness. That is life.”
-Captain Picard to Data, Star Trek: The Next Generation, ”Peak Performance”

Preface

This dissertation is submitted for the partial fulfilment of the degree of doctor of philosophy. It is based on work carried out between 2010 and 2014 in the Systems and Synthetic Biology group, Department of Chemical and Biological Engineering, Chalmers University of Technology under the supervision of Professor Jens Nielsen. The research was funded by the Knut and Alice Walennberg Foundation, BILS: Bioinformatics Services to Swedish Life Science and Chalmers Foundation.

Natapol Pornputtapong
March 2014

Abbreviations

ACID - Atomicity, Consistency, Isolation, Durability
API - Application Programming Interface
Corgi - C++ object-oriented graph interface
CRUD - Create, Read, Update and Delete
Dactyls - Derived ActiveRecord data model and query library for systems biology
DBMS - database management system
GEM - genome scale metabolic model
HMA - The Human Metabolic Atlas
Hreed - Human reaction entities database in object-oriented graph database
INIT - Integrative Network Inference for Tissue
JSON - JavaScript Object Notation
KEGG - Kyoto Encyclopedia of Genes and Genomes
KGML - KEGG Markup Language
MJSON - Map JavaScript Object Notation
Ondine - Omics navigator for data integration and evaluation
OODM - Object-oriented data model
OOGDM - Object-oriented graph data model
SBML - Systems Biology Markup Language
SVG - Scalable Vector Graphics
Sylfy - Systems biology library for ruby

1 Introduction

1.1 Background

Living organisms are complex systems composed of sophisticated and dynamic interactions involving a very large number of cellular components. To gain insight into such complex systems in an organism with the limited technology available, we typically consider them apart and individually study the specific properties of very small parts of the system. High throughput technologies, such as DNA sequencing, microarray, RNA sequencing, etc., has provided us with a great opportunity to observe the comprehensive variables of the system simultaneously with the development of system-level science of biology, i.e. systems biology (Kitano 2002). Systems biology aims to understand complex biological systems by integrating the information from the system components and their relationships to reconstruct computational models as a system scaffold from various data sources and levels such as genome, transcriptome, proteome, metabolome, interactome or reactome (Ideker et al. 2001).

Knowledge about human metabolism is important for the understanding of diseases, its diagnostics and for finding novel treatments. In order to understand human metabolism, human genome scale metabolic models (GEMs) as generic scaffolds of human metabolism have been built with examples such as the Edinburgh Human Metabolic Network (Ma et al. 2007) and Recon (Duarte et al. 2007). However, human metabolism is very complex and specific for each cell type. A generic model alone is not sufficient for a deep and full understanding of the whole human body metabolism. Several model reconstruction methods have been built based on different algorithms, but they have the same requirements, which are various layers and large amounts of data (Wang et al. 2012; Jerby et al. 2010).

With the emergence of high throughput technologies, the data has been expanded into many aspects including a number of data, data type and data collection 1.1. Consequently, databases such as GenBank (Benson et al. 2013), UniProt (The UniProt Consortium 2013), SGD (4), HMDB (Wishart, Knox, et al. 2009), KEGG (Kanehisa et al. 2012) and GEO (Barrett et al. 2011) have been developed and are available for the public. In order to utilize varieties of data sources, a systematic data processing pipeline, which includes data integration, organization and visualization respectively, is needed as shown in Figure 1.2. This remains a significant technical challenge in the field of computational systems biology.

The challenges of data integration consist of many aspects that usually occur during data propagation. 1) Database variation: most databases are developed by different aspects and provide different formats of data query interfaces. The information of data units from different data sources are sometimes not the same even if they describe the same thing, which is due to the aspect of data collection. For example, 2 databases describing certain protein characteristics from different aspects; one from the experimental aspect and the other from computational prediction, are not comparable. Moreover, with the heterogeneity of the query interface, most databases use the standard database

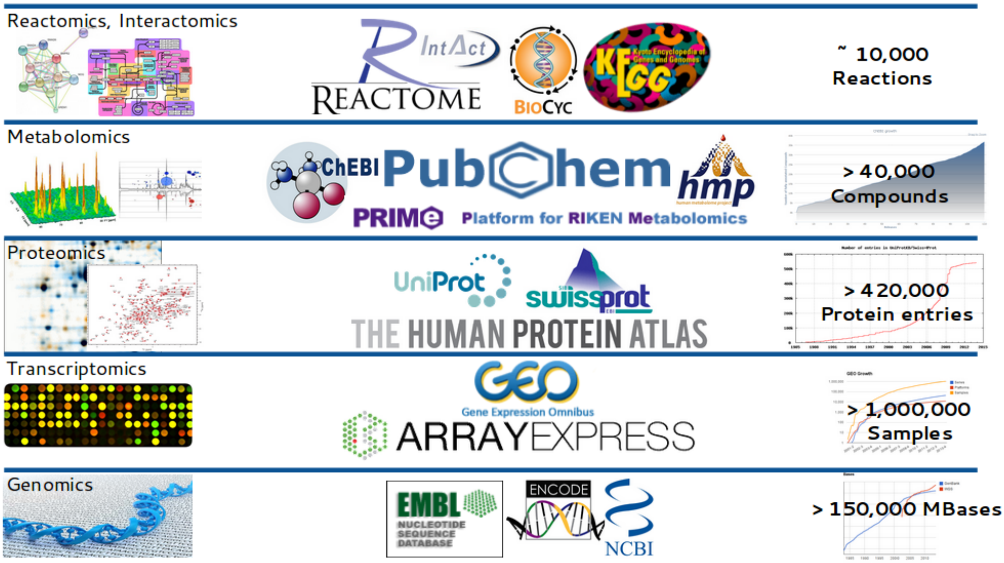


Figure 1.1 *Examples of public databases that are usually used in systems biology in three dimensional formats including data layers, number of databases in each layer and estimated size of the data.*

management system (DBMS) and implement widely-understood data query systems which enables the user to retrieve the right information easily, while some do not. Other data are provided in flat file with some interchangeable formats which requires certain libraries or software to read and interpret it. These barriers lead to problems concerning data quality. 2) The query: The complexity of the query from the databases increases with the complexity of the research question. This challenge becomes more critical when dealing with research questions in systems biology. 3) Data updates: Normally, data are updated locally. Most databases lack an interface to update data globally. Should there be one, it still remains a laborious task to update data from external sources into integrated data (Davidson 1995). All in all, these three factors make data integration and further management extremely challenging and error-prone.

The integrated data needs to be systematically organized, not only to ensure data integrity, but also to provide some interfaces to use the data. To build and implement a database for systems biological data is the same as a database system for other types of data. However, there are two problems that rarely exist in other types of database implementation.

1. Data aspect evolution: Data aspect, the way that we view and interpret the data, is implemented tightly in database design and can be changed by new knowledge discovery. When the data aspect changes, in most cases, the database schema has to be considered for redesign and reimplementation. Biological data is considered to be

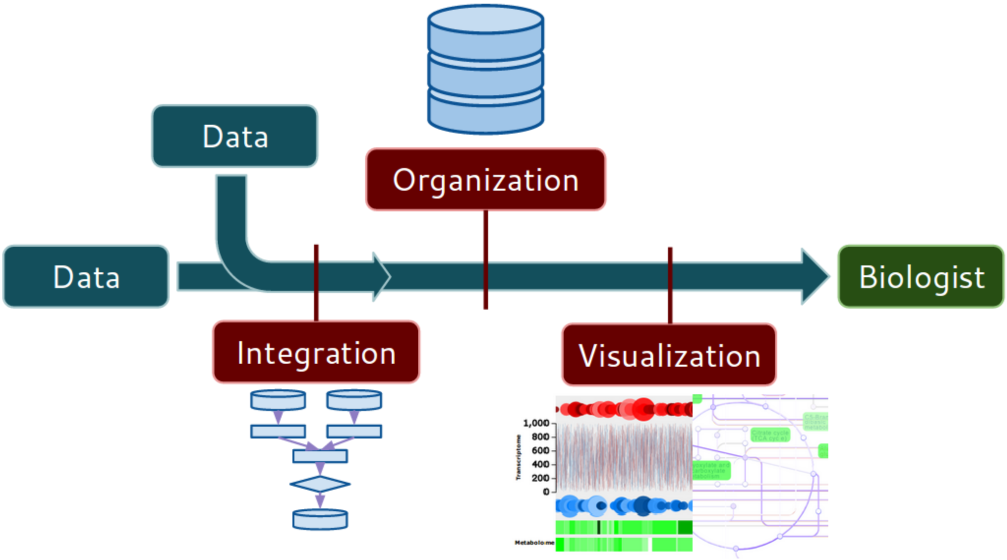


Figure 1.2 *Data processing pipeline; integration, organization and visualization.*

in the most complex data category that is constantly changing. These aforementioned reasons pose a severe problem for biological database implementation (Birney and Clamp 2004; Ozsoyoglu et al. 2006), which can be complemented by an adaptable data model design concept Millard et al. 2011.

2. Peopleware: Biological databases are very special and specific, in terms of data and utilization. Biological data that are usually stored in databases are in terms or words that can be ambiguous depending on the data aspect. Furthermore, the relationships of the information in biology are very complex, which also affects how data is stored and queried. Therefore, for efficiently managing this kind of data, a specific programming interface needs to be developed in parallel with the database design for propagating the data and utilizing it. This means that developing and implementing a biological database is not a database design problem, but it is associated with biological and programming problems (Birney and Clamp 2004), which means that there is a need for people who have multidisciplinary knowledge.

Besides querying data, visualizing data is another easier way for the user to investigate and explore the information stored in the database. Visualization is helpful for several research fields in biology, but with the increasing complexity of biological data particularly in systems biology, building a visualization system to illustrate all the information of the whole complex system remains challenging (Tao et al. 2004).

This thesis focuses on database and visualization system development integrated by web-based technology as a comprehensive research platform to support GEM reconstruction and distribution.

The first version of the database, called Human Metabolic Reaction, was built relying on relational database model with SQL-based database management system to support INIT algorithm (Agren et al. 2012), for which GEM reconstruction is based on as described in **Paper III**. However, due to the complexity of the data used in the modeling work, there were too many data tables used in the data model causing less efficiency of the querying process and resulting in a loss of accuracy in the queried data. To overcome this problem, a database management system was newly developed relying on a combination of object-oriented and graph data model in programmatic data layer (conceptual data layer) and a document-oriented data model in actual data storage layer (physical data layer). Whole database processes, including **C**reate, **R**ead, **U**pdate and **D**elete (CRUD), are provided in the database API library base, which was initially developed in C++ programming language named **Corgi** (**C**++ **o**bject-o**r**iented **g**raph **i**nterface) with more efficiency in speed and memory optimization (**Paper II**). However due to the nature of the C++ language, this library is quite difficult to use by biologists. To provide an easier alternative for biologists to use the database system, the second database API library was built in the Ruby scripting programming language called **Dactyls** (Derived ActiveRecord data model and query library for systems biology). This library provides classes and functions to support data modeling and general database activities and particularly the new data query system mimicking the biological processes, which biologists can use without any effort (**Paper IV**).

To make high throughput data easier to digest for biologists, an effective visualization system is needed. Ondine (Omics navigator for data integration and evaluation) was developed to interactively visualize a multi-level omics data, which allow users to simultaneously navigate transcriptomic, proteomic and metabolomic data on biological pathways like the KEGG pathway using Ruby on Rails and JavaScript (**Paper I**). All these software developments are intended to support data expansion and utilization not only for the Human Metabolic Atlas project, but also hopefully for general uses in the research community (**Paper IV**).

1.2 Thesis structure

This thesis represents a summary of 4 publications and is divided into six chapters. Chapter 2 describes the database design for systems biology data, which is a part of **Papers II**, **III** and **IV**. Chapter 3 describes the database management system API library development, pertaining to **Papers II** and **IV**. Chapter 4 describes the visualization of multi-omic data on metabolic pathway maps from **Paper I**. Chapter 5 details about the development and implementation of the Human Metabolic Atlas from **Paper IV**. Chapter 6 summarizes the perspective of this work.

2 Database design and implementation for systems biology data

The first point to consider before using a database is the data model, which is a scaffold of data structure used by the database management system when the data is stored and queried. Compatibility of the data model to the data can affect the efficiency of the database system and also the integrity of the information inside the database. This chapter describes briefly the concepts and methods to design data structure that was used in **Papers II**, **II** and **IV**.

2.1 Database design concept

The key benefit of using a DBMS is that data are safely and accurately shared to restricted users or even to the public. To achieve such goals, the basic concepts of a DBMS, including the ACID (**A**tomicity, **C**onsistency, **I**solation and **D**urability) properties, are to be considered in the database design. To control the validity of data changes occurring when the user performs updates to the database, the atomicity concept is applied. In particular, only successful transactions will be committed to the database, otherwise nothing will be committed. Consistency ensures control of data integrity when multiple users are working at the same time. The isolation concept is used for preventing interference between two transactions working on the same data object. The last concept considered is durability, which ensures that the committed data will never be lost (Barry 1996). The design of the data structure follows a ANSI/X3/SPARC proposed data architecture, which uniquely separates the view of the data structure into three layers (Steel (jr.) 1975):

1. An external layer, which is the first layer, could be considered as the outer layers of the database of data abstraction in the database system. It represents the entities of data to users or applications.
2. A conceptual layer, which is the second data abstraction layer, represents the entities of data that are assembled from the physical layer and can be transformed to the external layer as needed.
3. A physical layer represents the concrete data structure that is implemented in an actual file system and is only used by the DBMS.

All of these three layers were set up independently. There are several data models that can be used in database design. Table 2.1 compares all of the data models that were used in this thesis.

2.2 Data identifiers

To make data more consistent, identifiable, understandable and exchangeable, several data identifiers were implemented for chemical compounds, cross references and annotation

Table 2.1 Comparison of data model terms and concepts based on the relational-table data model.

Relational-table	Object-oriented	Graph	Document-oriented
table	class, type	-	collection
record, row	object	node, edge	document
field,column	attribute	attribute	attribute,field
SQL	OQL	-	-
table join	relation aggregation	graph transversal	link

words during data propagation in the integration processes. For reactions, a specific identifier was newly developed as described below.

2.2.1 InChI and InChIKey

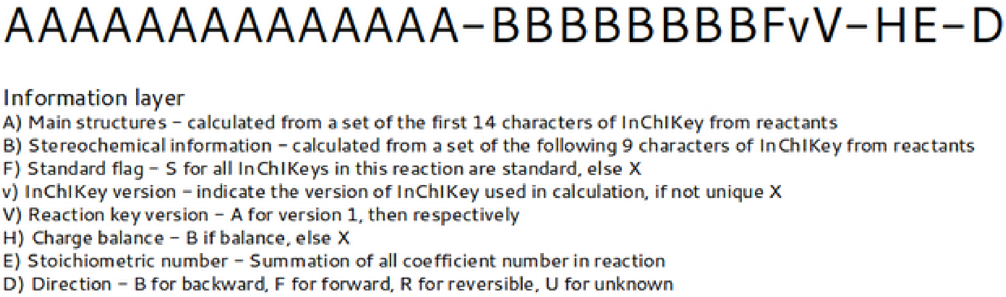
To provide unique identifiers for chemical compounds instead of using the conventional and ambiguous identifying names, IUPAC developed a unique computer readable identifier of chemical compounds named InChI and InChIKey. InChI is comprised of several data layers that are specifically generated from a molecular structure diagram (Heller et al. 2013). While the length of InChI increases by the number of atoms in the molecule, which is not suitable for use as database identifiers and is also unreliable to be used as search keywords, the length of InChIKey is constant. InChIKey contains 25 characters of encrypted InChI string by SHA-256 function is comprised of 5 informative layers. The full description of InChI and InChIKey can be obtained from Heller et al. 2013; Williams 2012; Pletnev et al. 2012; Bachrach 2012. InChI and InChIKey are minimum requirements for every SmallMolecule objects to be put into the database.

2.2.2 Miriam

One criteria of database design is to support data integration from public databases. Each database uses its own identifying system, which usually contains only numbers. To avoid the ambiguity of cross reference identification, the Miriam (Minimum Information Required in the Annotation of Models) registry was applied in the conceptual data layer. To provide unique identifiers regardless of the actual data source, the Miriam registry was provided as an URN (Unified Resources Name) string which comprised of three parts. The prefix is always ‘urn:miriam’ to specify the register source that is from Miriam, followed by the namespace of the data collection source. The last part is the identifier itself. Miriam registry also maintains the actual location of data, which can be useful for avoiding dead links (Juty et al. 2012).

2.2.3 Reaction key

The key bottleneck of reaction data integration is the inability to compare data among different sources of database due to incompatibilities of data representation such as the reaction description format and reactant identifiers. A unique identifier for each reaction



AAAAAAAAAAAAAA-BBBBBBBBFvV-HE-D

Information layer

A) Main structures – calculated from a set of the first 14 characters of InChIKey from reactants

B) Stereochemical information – calculated from a set of the following 9 characters of InChIKey from reactants

F) Standard flag – S for all InChIKeys in this reaction are standard, else X

v) InChIKey version – indicate the version of InChIKey used in calculation, if not unique X

V) Reaction key version – A for version 1, then respectively

H) Charge balance – B if balance, else X

E) Stoichiometric number – Summation of all coefficient number in reaction

D) Direction – B for backward, F for forward, R for reversible, U for unknown

Figure 2.1 *Information layers of reaction key.*

is needed. The international chemical identifier for reactions (RinChI) was developed by (Grethe et al. 2013). RinChI is mostly analogous to InChI in providing chemical structure information of reactants and directionality of reactions. However, RinChI is still lagging in some information, which is usually considered in GEM reconstruction; such as stoichiometry and charge balance (Kumar et al. 2012). The reaction key was hereby developed to provide a more comprehensive description of a reaction in a constant length string. The reaction key string is comprised of 5 essential informative layers of the reaction; the structure connectivity and stereochemical information of the reactants, charge balance, stoichiometric number and reaction direction and 3 version control layers for identifying the version of this reaction key as shown in Figure 2.1. In order to generate a reaction key, two groups of molecules, the substrate and product, are described as lists of InChIKey. The list of substrates and products are sorted within a group and again among the groups. The structure connectivity layers from each InChIKey in the sorted list are concatenated prior to calculating the structure connectivity layer of the reaction key using the SHA-256 function. The stereochemical layer of the reaction key is calculated in the same way as in the structure connectivity layer but by using the stereochemical layers from the sorted list. The charge balance is the difference between substrate and product charge summation, which is calculated by using the charge layer in InChIKey. The stoichiometric number is the summation of reactant coefficients described in the reaction. The reaction key is used in the database as the main identifier of the reaction in this database design.

2.3 Data integration

Data integration is a key step and also the most tedious task in the data utilization process. There are three conceptual tasks that can be applied.

1. Data model transformation and semantic schema matching: the underlying data schema has to be converted to a common schema before integrated. The transformation needs to preserve all the relevant information, not just the data itself, which means that the common schema has to reach the criteria. In order to construct the global schema, the data schema from each data source needs to be matched together

to find corresponding or conflicting schema components. This is an important problem during semantic integration. Some components are named in the same way, but refer to different things depending on the database development aspect. Recently, there are no automatic processes available that can be used in semantic matching because of the ambiguity of the schema naming, which is uninterpretable by the computer.

2. Schema integration. A global schema is constructed relying on the matched schema, for which there are 2 approaches, top-down and bottom up. The top-down approach is to build the common schema first and then continue mapping with the schema of the data source. On the other hand, the bottom-up approach is to build the common schema explicitly from the matched schema of data source. The global schema is generally a union set of data source schema components, which depends on the degree of integration. With tight degree of schema integration, schema components of the data sources are mapped precisely to the global schema, after all the conflicting schema components of the data sources have been solved in order to ensure data accuracy. Whereas, loose degree of schema integration is simply just to pool all schema components of the data sources together in order to build the global schema.

3. Data transformation and data matching. The transformation of the schema level to the global schema can possibly affect the interpretation of data in the underlying data sources. The data from the data sources need to be transformed correspondingly to the global schema in order to initiate data matching. The degree of data integration depends on how the integrated data is collected. Materialized degree of data integration is a physical based data propagation and maintenance that information from all data sources are actually collected and maintained by the target database. The advantage of this solution is the performance of the target database system, but it is costly in terms of resources and time to maintain and update the data. Although the view degree of data integration, which can be considered as virtual data propagation, is just collecting the hyperlinks to data from data sources, target databases do not have to maintain the whole data. However, the efficiency and accuracy in querying across the database depends on the communication between the database and data structure of the data sources.

From schema and data integration perspectives, the degree of integration can be classified into two dimensions; tight vs. loose for schema integration and materialized vs. view for data integration. Tight and materialized degree of integration requires intensive cost of implementation in order to particularly match the schema of all database sources and propagate the complete information from each database together into a database, but provides a high integrity of data. Although, the loose and view degree of integration requires less effort to transform data schemata and to collect the data, however, they provide less data integrity. By all means, choosing the degree of integration is to choose between the cost of implementation and efficiency.

In this work, the database design was aimed to be a global data schema that can be used in data integration. The integration algorithm was not implemented in database

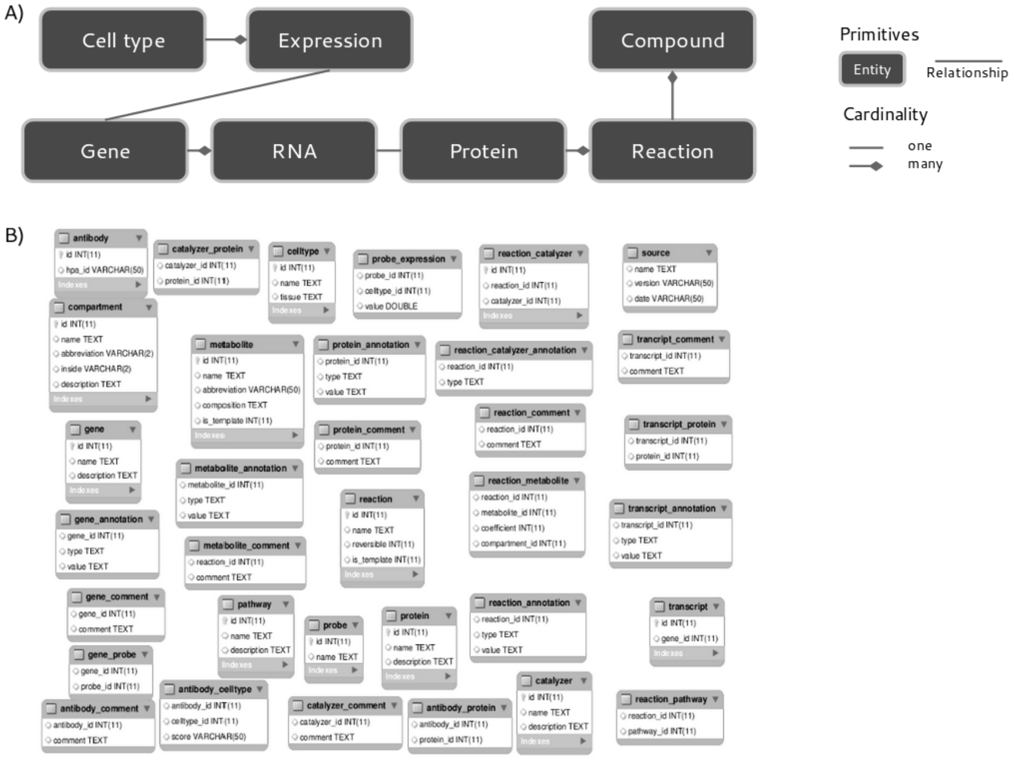


Figure 2.2 *A) Conceptual data structure of the HMR database with boxes representing the data components and lines representing their relationship with cardinality. B) The HMR database was built using the SQL database. Conceptual data structure was converted to relational data tables.*

API, but some of the functions were implemented to support data integration conceptual tasks, including file format parsers and web service retrieval function.

2.4 Database design and implementation of HMR database

In order to provide reliable data for GEM reconstruction, a database called HMR database has been developed as described in **Paper III**. This database was basically designed to populate the metabolic network information and to provide a platform for further omic data integration especially expression data. A MySQL database management system was used for managing this database. The relational table data model, which manages a collection of data entities in a table and each data as a tuple, was used in this database design. The data schema has a hierarchical structure as shown in Figure 2.2 A. After data normalization, the schema has been converted to the tables as shown in Figure 2.2 B.

Table 2.2 Version and provided data of data sources that were used for the HMR database construction. (Adapted from **Paper III**)

Database	Data category	Source format	Version
Recon1	Reaction, metabolite	SBML	Jan 31, 2008
EHMN	Reaction, metabolite	Excel	June 6, 2009
HumanCyc	Reaction, metabolite	Text file	12.5
KEGG	Reaction, metabolite	Text file	48
HPA	Proteome	Text file	7.1
HMDB	Metabolome	Text file	2.4
BioGPS	Transcriptome	Text file	2.0
HepatoNet1	Reaction, metabolite	Text file	March 1, 2011

Table 2.3 Number of stored data in HMR by category. The numbers in parenthesis are numbers of unique data. (Adapted from **Paper III**)

Data category	Number
Gene	2,366
Compound	9,581 (3,547)
Reaction	9,922 (6,319)
Compartment	8

In order to support GEM reconstruction, the HMR database was built for propagating major information from the existing human genome scale metabolic models, Recon1 and EHMN, as well as integrating the required information for reconstructing algorithms from external data sources including HumanCyc, KEGG, HPA and HMDB as shown in Table 2.2. Data was integrated in the tight-materialized degree of integration, which means that all information were transformed and standardized. To reduce the ambiguity of the metabolite and reaction information, the InChI and KEGG identifiers, which are unique for each chemical structure, were used for data standardization. Metabolites with lacking identifiers were not propagated into the database, as well as their corresponding reactions. Each reaction was assigned to one or several compartments relying on the information available from the existing models. Without prior information from the models, localization of each reaction was inferred from HPA first, thereafter Swissprot and GO respectively. After the data propagating process, the total number of data is shown in Table 2.3.

2.5 Database design and implementation in Hreed

In **Papers II** and **IV**, the data structure design of Hreed, a database for human reactions and related omic data representing an ongoing endeavor to serve researchers in human metabolic network data analysis and GEM reconstruction, was described. It is a specific data model that was designed and developed by taking the following into account: 1) the ability to integrate multi-level omics data; 2) that biological data are complex, heterogeneous, and dynamic (Ozsoyoglu et al. 2006); 3) the diversities of resources in terms of data model, semantic heterogeneity, data completeness and data correctness; 4) reusability, extensibility, flexibility and interoperability of the system; and 5) integrity, consistency and reliability of the data in the database.

2.5.1 Conceptual data layer

This abstraction layer serves as an interchangeable data structure between the user and database management system. All database processes and activities were implemented relying on this data layer. Database design usually begins here.

object-oriented data model (OODM)

In an object-oriented data model, a real world object is represented as a data object, which can be distinctly identified. This concept is applicable for biological information, which is apparently heterogeneous and sophisticated (Okayama et al. 1998). A data object is characterized by their class or object type and its attributes (Zhao and Roberts 1988). A class or object type is an entity type that has a well-defined state, identity and behavior in the application domain and should represent a tangible and visible entity type (Hoffer et al. 2011). Attributes are normally described as attribute-value pairs, which are comparable to fields and their data in the relational-table database. In the OODM, there are three types of attributes (Zhao and Roberts 1988):

- Value attribute is a primitive data type, which can includes boolean, integer, floating point and character or string.
- Group attribute is a group of primitive data type. In database implementation, this is normally described as an array or set.
- Aggregation attribute is a special attribute type that is used for referring to another object normally by Object ID (OID).

Relationship among objects can be easily specified using the aggregation attribute. However, to aggregate objects together, it is required to clearly specify about the type of object that can be aggregated within the class during the first implementation to preserve data integrity. This is not suitable for biological data, which are very dynamic. The OODM was implemented in this database design without aggregation attribute type.

graph data model

Graph is a collection of nodes connected by edges. To apply graphs into the data model, a data entity is described as a node with node attributes, which is the same as in OODM. Nodes are connected by edges to illustrate their binary relationship. Edges are typed and always stored pointers to start and end nodes. This data model is suitable for storing less descriptive data entities with complex relationships.

object-oriented graph data model (OOGMD)

With some limitations of these database models, a data model was designed by using a combination of the OODM and graph data model in programmatic data layer (conceptual data layer) storing tangible biological entities as objects and their relationships in binary relationship of the graph model to fit with the high complexity data that are used in GEM reconstruction and multi-omic data integration.

In this work, classes were designed by adapting the class description in the BioPAX ontology. BioPAX is a standard language used to define biological pathways including related entities to support the development of databases and computational tools. A set of well-defined abstraction classes, covering all real world phenomena used by systems biologists, were implemented in a language standard (Demir et al. 2010). The implemented classes were strictly designed based on object-oriented programming concepts: 1) Data abstraction and encapsulation, properties and data structure of a class have to be protected from procedures and users. Data in the classes can only be accessed through the class interface called operator. This design concept is used to control data consistency inside the object; and 2) inheritance is a mechanism of code reusability, which is a powerful object-oriented approach. Subclasses were generalized from its ancestor class called superclass and inherited the basis of class definition; such as attributes and methods from their superclass Bertino and Martino 1993. This concept makes the library more modular and easy to organize. To represent the data models, a specific tool called Unified Modeling Language (UML) Object Management Group 2011 was used to represent the static view of the conceptual data structure. Biological components; transcripts, proteins, compounds and reactions are defined as physical entity class mimicking node in the graph data model. Whereas, relationships among the classes are represented separately as relation class mimicking edge in the graph data model.

2.5.2 Sub-conceptual data layer

To avoid deep technical development of the physical data layer, the conceptual data layer was implemented as a separated API on top of the conceptual data layer of the underlying database management system named MongoDB, which can be considered as a sub conceptual data layer. MongoDB was chosen because of its conceptual data layer, which relies on the document-oriented data model. This data model is a dynamic schema and a fluent polymorphism data model, which perfectly supports the data schema changes. This layer of data structure is considered to be an interchangeable data model between the conceptual data layer in the database API and the physical data layer in MongoDB. In Sub-conceptual data layer, data entities are considered as documents, which are gathered together in a collection. Data objects are managed as documents in the BSON format, which is a binary version of the JSON documents. The JSON document is a text-based document standard that was designed for human-readable data interchange. The structure of the documents was derived from the JavaScript language for representing associative arrays with a set of attribute-value pairs. Each attribute in a data object was converted to attribute-value pairs. Data and relation objects were stored separately in different collections.

Table 2.4 Summary of the Hreed database. (Adapted from **Paper IV**)

Datasets	Source	Data types	Imported/Total
Ensembl gene 69	biomart.org	Gene (and Chromosome)	62311
Ensembl transcript 69	biomart.org	Transcript	213272
UniProt 2012_09	uniprot.org	Protein	19084
HMR compound	metabolicatlas.org	Compound	1692/3539
Pooled dataset	metabolicatlas.org	Compound	72594
HMR reaction	metabolicatlas.org	Biological reaction	5282/5526

2.5.3 External data layer

The external data schema is not different from the schema in the conceptual layer. However, in order to control data consistency during requests and data updates from applications, the data access library was designed to support general operations that are requested from applications or clients such as by querying, inserting, updating and deleting data with the class and object method. These operations were developed by using database operators from the MongoDB library and implemented using some build-in processes to control the integrity, consistency and reliability of data following the ACID properties. These operators were defined specifically for each type of data class. The output data objects returned from the operators were instances of data classes. The actual data structure was encapsulated in the class.

To provide feasible data query interface to users, particularly biologists, query methods were designed based on their actual object behaviors. Instead of using general searching query languages like SQL, where users have to know the exact conceptual data structure, specific methods named as their tangible object behavior were implemented; such as finding the transcript's object from gene by using the 'transcribe' function or proteins from transcript by using the 'translate' function, etc. The full detail of the implementation will be illustrated in Chapter 3.

2.5.4 Implementation of Hreed database

The Hreed database was developed to collect reaction, metabolite and gene-reaction relationship information of models deposited in the HMA repository for further GEM reconstruction and data analysis. Several kinds of data including human gene, transcript, protein, small molecule and reaction data, as shown in Table 2.4, were transformed into data objects as described in the conceptual data structure and then propagated into the database. Small molecules and reactions data, regarded as the major information in Hreed, were populated from HMR1.0 (Mardinoglu et al. 2013) with minimum data requirement concerns according to one of the design criteria, data integrity. The minimum requirements for propagating a metabolite are InChI and InChIKey and for the reaction data, a reaction key, whose calculator was provided by the Sylfy library. To support reaction data expansion and to assist users when adding new reactions to the database in the future, small molecule data was incorporated from external compound databases including HMDB (Wishart, Tzur, et al. 2007), LMSD (Sud et al. 2007), ChEBI (Hastings et al. 2013) and PubChem (Bolton et al. 2008) with full InChI annotation.

Based on the database design, the database schema was implemented with a tight degree of integration, in which all of data source schemata had to be transformed and were precisely matched with the conceptual data structure before data transfer. This approach is time consuming, but ensures high quality of the data integrity.

3 Database management system API library development

3.1 Corgi (C++ object-oriented graph interface) API library (Paper III)

This chapter describes about the development of Corgi, a database API for systems biology data that represents an ongoing endeavor to serve researchers in systems biology and to provide alternative solutions for vital issues in data handling, access and integration. It is a specific database API that was designed and developed by taking the following into account: 1) the ability to integrate multilevel omic data; 2) that biological data are complex, heterogeneous, and dynamic (Ozsoyoglu et al. 2006); 3) the diversities of resources in terms of data model, semantic heterogeneity, data completeness and data correctness; 4) reusability, extensibility and interoperability of the system; and 5) integrity, consistency and reliability of the data in the database. An object-oriented concept was adopted for the design of the database schema, which represents practical information as an object with related attributes and a variety of relationships. This concept is applicable for biological information, which is apparently heterogeneous and sophisticated (Okayama et al. 1998). Corgi was developed in C++ and included a library providing important functions to manage and interact with the system.

3.1.1 Global System Architecture

Corgi is a specialized database API developed using the C++ programming language based on conceptual data structure of the database design. The overview of the system architecture is shown in Figure 3.1. As the base of the system, the physical layer is managed by a document-based management system, MongoDB, which contains the necessary interfaces such as an interactive shell and web services. However, MongoDB is not designed to manage structured data. This may cause problems in data integrity, consistency and reliability. Corgi was therefore implemented as a database API, providing vital functions to manage transactions between developers and the system making it easy to populate and transform data.

3.1.2 Library architecture

The Corgi API library was developed in C++ on top of the MongoDB driver. It provides 4 class collections; data wrapper, database, parser and services.

3.1.3 Data wrapper class collection

To manage complex data in systems biology, a specific OOGDM was implemented in the data wrapper class collection. This class collection was developed in an object-oriented approach to manage the data structure at the conceptual layer and to work as

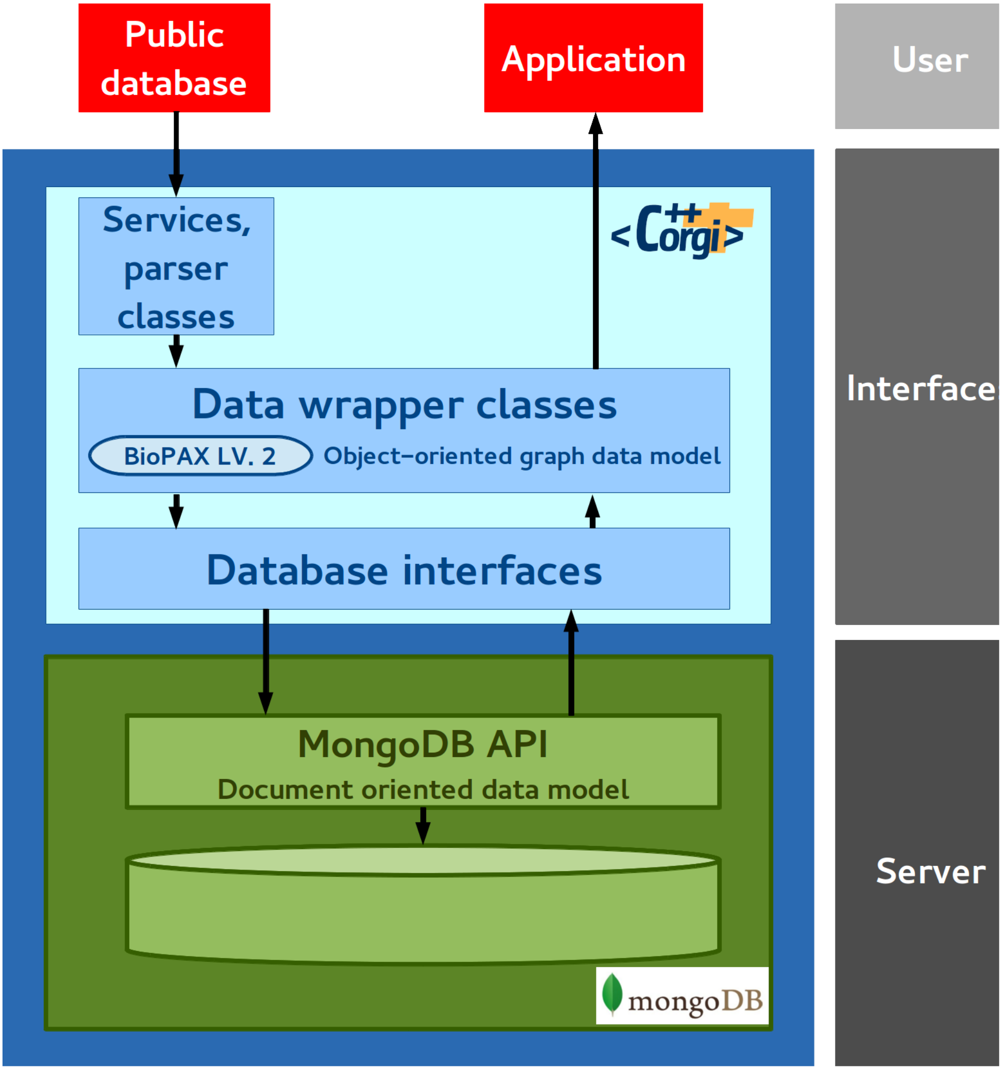


Figure 3.1 *The Corgi API architecture. The Core library is comprised of data wrappers, services and data parser classes built on top of the MongoDB database management system. (Adapted from* ***Paper II****)*

an interface between the developer and the database by hiding the actual implementation of the physical data structure. The design of the class structure relied on the basis that controls data integrity and consistency. The wrapper class structure was adapted from the ontology class of Biological Pathway Exchange (BioPAX) (Demir et al. 2010). All classes in the data wrapper class collection were specialized from the superclass ”Object”. The BioObject class and its subclasses represent the major type of data that can be populated into the system. The PhysicalEntity sub-classes support molecular entities including small molecules, DNA molecules, genes, RNA molecules, proteins and molecular complex data. The Interaction subclasses support biological reactions, molecular interactions, genetic interactions and control interactions. There are relationships among the concrete classes following the biological relationships of real biological objects to support the data integration of multilevel data into the network as shown in Figure 2b. The Literal subclasses were designed to support the general data structure that is commonly used in the BioObject derived classes, and the Literal subclasses, except for the Relation class, were not placed in the physical layer of the database system as an independent document. In the conceptual layer, the Relation class was included in the BioObject derived class as object data members, however, when objects were inserted into the physical layer, the Relation class instances were placed separately from BioObject derived classes. Subclass in Corgi does not support sub-typing ability. Users cannot query or refer to the sub-class by their superclass. Figure 3.2 is an illustrated overview of the implemented classes in the library.

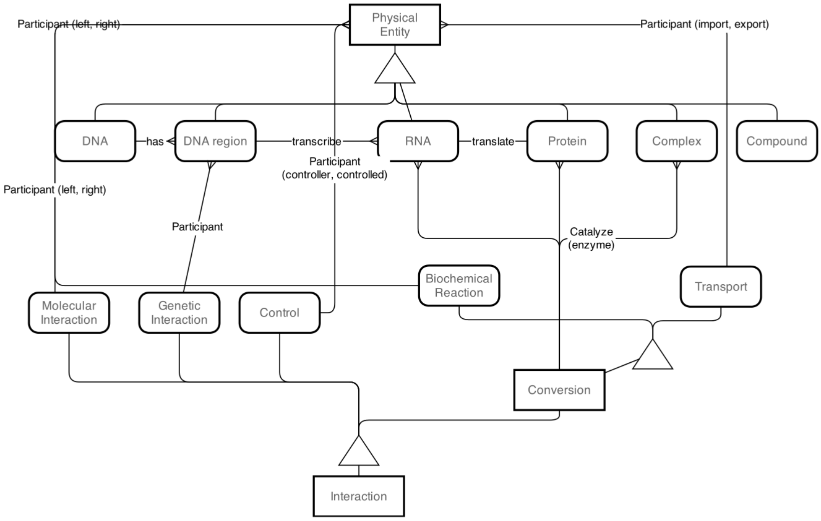


Figure 3.2 *The object derived classes and their relationships. A) Class diagram of the object derived class. The boxes represent the classes. The diagram illustrates the relationship among the classes. (Adapted from **Paper II**)*

Parser class collection

There are several public databases where users can retrieve and dump the data into their own database. Unfortunately, each database provides data in different formats without any proper parser libraries in C++. To cope with this issue and support data integration concepts, the library provides classes for parsing general data, which are usually used in systems biology. In particular, the data files will be parsed into a specific instance with available interfaces for database managers to extract the right information.

Because of the inconsistency of data formats among the public databases, the library does not automatically format the data to the system data structure. An additional program is required to extract the desired information from the parser class interfaces and then transform the data into the system data structure by using the data wrapper class collection. The parser class collection supports general formats used in most biological databases including BioPAX, PSIMI, ChEBI OWL and OBO OWL format. The BioPAX parser is based on Level 2, Release Version 1.0 Cary et al. 2005. Codes from XMLParser (version 2.42 of Business-Insight International under the AFPL licenses) were implemented in all parser classes.

Service class collection

The ”REST” class was developed using the cURL library [http://curl.haxx.se] to send request- packages to the REST (Representational State Transfer) server and retrieve responses back. Only the simple GET and POST methods, supported by most public web services, were implemented in this class. The responses are commonly returned in the XML format that is parsed automatically by an XML Parser. There were two service interfaces included in this class collection, MIRIAM registry and BioPortal services. The MIRIAM registry service was implemented in the database system as a standard for cross-reference identities (ids). Some of the MIRIAM web service interfaces were also implemented in this class collection by using the ”REST” class, which are useful in resolving the MIRIAM registry to the original URL and back to the reference. BioPortal (Whetzel and Noy 2011) is a web portal of the National Center for Biomedical Ontology (NCBO) for providing information about biomedical ontology terms and to map between terms in different ontologies. All terms in the database system are forced to use only registered ontology terms. To make it simple for developers, interfaces for the BioPortal RESTful web service were implemented in the services class collection to request the ontology term information and to map the term id.

3.1.4 Sub-conceptual layer data structure

The objects from the PhysicalEntity and Interaction subclasses were stored in the database as JavaScript Object Notation (JSON) documents (Ecma 2009). The management system implements the JSON documents in a binary version called BSON. All documents are pooled together in the ”object” collection without a relation field. The system recognizes the class of an object from the term in the ”type” field, which is automatically defined by the system when the object is initiated. To improve the consistency of complex relationships, the graph database concept was implemented in the physical layer. Relationships among the objects are represented as edges of the graph and stored separately in a different collection named ”relation”. The relation objects refer to the related objects (i.e. a node of the graph) by id.

3.1.5 Database activities

In general, most of the noSQL DBMSs are lagging some ACID properties. As well as in MongoDB, it only provides the A-atomicity operation to the single document and

D-durability by using a journaling system. To maintain integrity, consistency and reliability of data within the database during the committing process and to ensure that most transactions follow the ACID principle, all activities in the database have to be done using the specific functions from the Corgi API library. Four basic functions were provided to manage general activities inside the database system as shown in Figure 3.3.

Create To avoid duplication, data insertion begins with comparing between the cross references from 'dataPrimarySource' and 'xref' field in the object and the index of the 'object' collection. If there is a match, the system will raise an error message to the user and terminate the insertion process. Afterwards, relation documents will be generated. If there is no error, an object document will be generated and inserted into the database followed by the atomic insertion of the relation documents. As previously mentioned, the database system does not support multiple document transactions across the collections. During this 'insert' operation, the system needs to insert both the 'object" document and related 'relation' documents. To maintain the ACID properties, if the atomic insertion of the relation documents results in an error, the object document added previously will be deleted. The insertion process will be terminated following this.

Read This function uses the query engine of the database system to obtain objects from a submitted query string. Related relationship documents will be queried and combined with the object document. This query function also provides a process to format the resulting documents into elements of the data wrapper class.

Update Update transactions start with masking the original document into a temporary one followed by inserting the updated object document into the database. If the insertion process fails, the original version of the document will be returned to the database. Otherwise, it will be removed.

Delete At the beginning, all objects are obtained from a query string provided by the users. Each 'object' document will be masked as a temporary document and subsequently, the related relationship documents will be removed from the 'relation' collection with an atomic transaction. At the end, the temporary document will be removed from the database. To maintain the ACID properties, if there is an error during the process of removing the relationship documents, the object document will be returned back to the database.

3.2 Dactyls (Derived ActiveRecord data model and query library for systems biology) API (Paper IV)

3.2.1 Global System Architecture

Dactyls is a database API developed based on the ruby scripting language relying on object-oriented and graph conceptual data structure of the database design. This API library

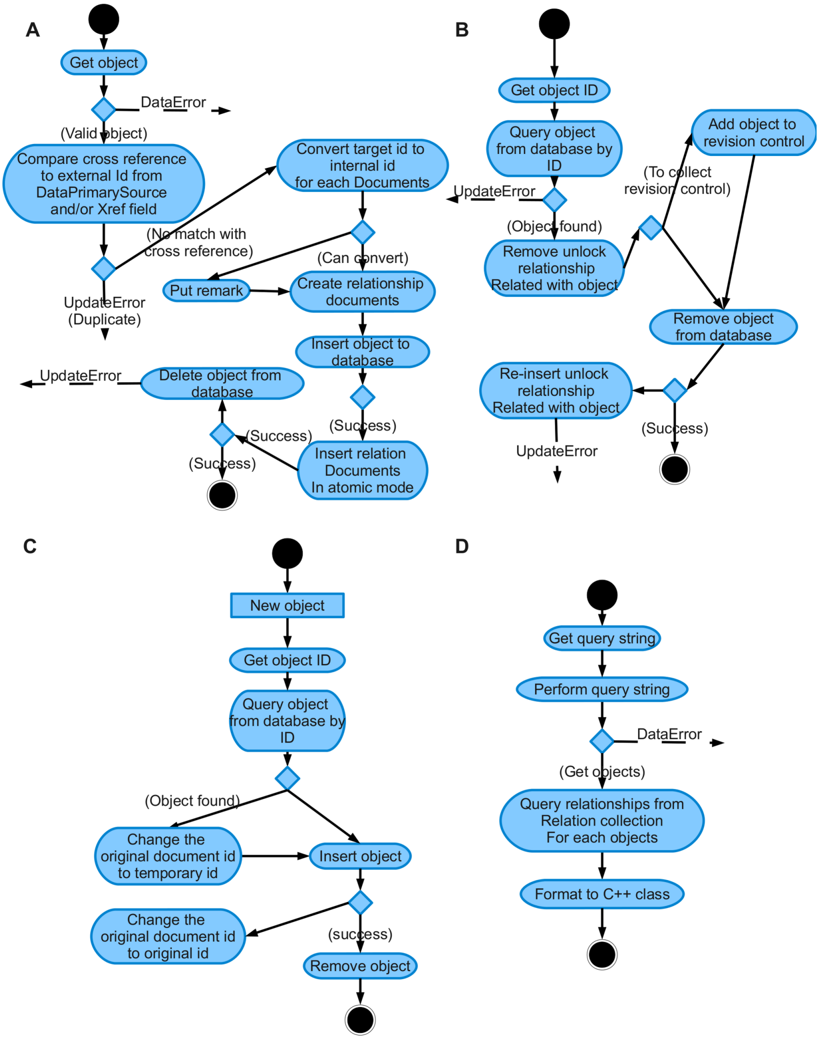

A
Get object
DataError
(Valid object)
Compare cross reference to external Id from DataPrimarySource and/or Xref field
(No match with cross reference)
UpdateError (Duplicate)
Convert target id to internal id for each Documents
UpdateError
(Can convert)
Put remark
Create relationship documents
Insert object to database
(Success)
Insert relation Documents In atomic mode
(Success)
Delete object from database
UpdateError
(Success)
B
Get object ID
Query object from database by ID
UpdateError
(Object found)
Remove unlock relationship Related with object
(To collect revision control)
Add object to revision control
Remove object from database
Re-insert unlock relationship Related with object
(Success)
UpdateError
C
New object
Get object ID
Query object from database by ID
(Object found)
Change the original document id to temporary id
Insert object
Change the original document id to original id
(success)
Remove object
D
Get query string
Perform query string
DataError
(Get objects)
Query relationships from Relation collection For each objects
Format to C++ class

Figure 3.3 *UML activity diagrams of the general database transactions. A) create B) delete C) update and D) read*

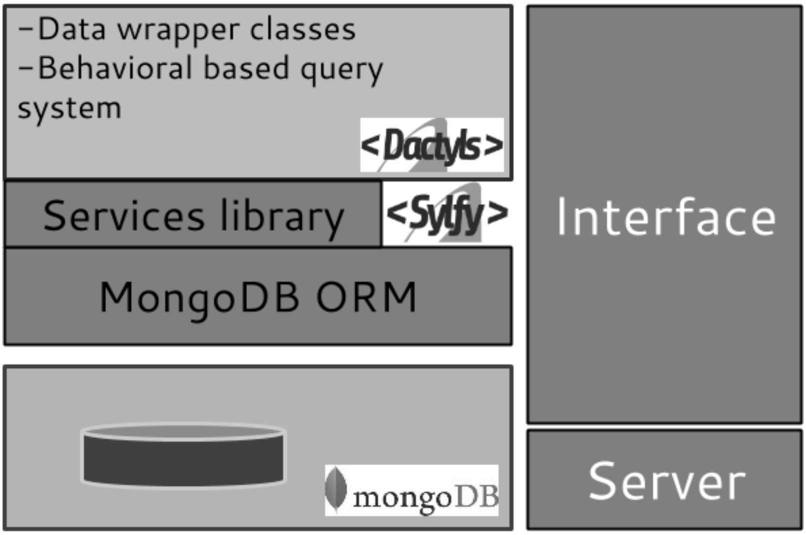


Figure 3.4 *The Dactyls API architecture. Dactyls was developed to serve the behavioral query system and accommodate further database expansion incorporated with the service library Sylfy and the ORM library on top of the MongoDB database management system.*

was developed in parallel to Corgi. While Corgi was especially developed for programmers and database system managers who have good knowledge about the C++ programming language, Dactyls is tailored for end users such as biologists. The overview of the system architecture is shown in Figure 3.4. This database API was also developed based on the object-oriented graph data model in the conceptual data layer using MongoDB as the underlying database system. Dactyls was therefore implemented as a database API, providing effortless query system for biologists and supporting general CRUD activities.

3.2.2 Library architecture

Dactyls was developed using the object-relational mapping (ORM) system, MongoModel, to convert data models between the document-oriented data model in the sub-conceptual data layer into the object-oriented graph data model in the conceptual data layer. ORM is a computational technique to convert incompatible data models, usually relational databases, into object-oriented models by creating virtual object-oriented databases as an interface between the database management system and application or API. MongoModel is a Ruby scripting language ORM system specific to MongoDB, used as a base DBMS, to provide several functions to create data objects and CRUD activities. With this ORM, Dactyls can fulfill the expansibility of design criteria, support the dynamic properties of biological data and support subtyping. Utility classes including web services, file format conversion, chemical data conversion and reaction key calculation were developed separately in the Sylfy library.

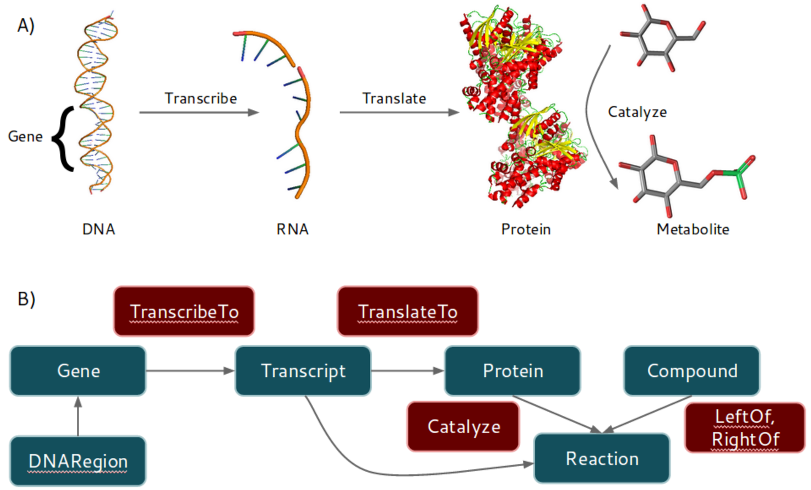


Figure 3.5 *Data models in the conceptual data layer. (The DNA, RNA and protein molecules were generated from PDB files of 2O61, 1PNS and 3HM8 respectively using Pymol (www.pymol.org). Glucose and Glucose-6-phosphate structures were obtained from PubChem (Evan et al. 2014). (A) Biological entity categories and their functions. (B) Data classes reflecting each biological entity category and relation classes for their functions.*

Data wrapper classes

Data wrapper classes are the implementation of the conceptual data layer following the object-oriented graph data model. With OODM, the classes were implemented imitating tangible biological components, genes, transcripts, proteins, compounds and reactions. All classes were derived from the 'Node' (represented as blue boxes) and 'RelatedTo' (represented as red boxes) classes for all the biological components and their relationships as shown respectively in Figure 3.5.

External data layer

The top data layer of the database provides the user's view of the data normally as in the query system. To assist biologists in querying data from the database, external data layer was developed using the behavioral modeling concept. The query functions were designed to retrieve related information by following real object behaviors. The data structure relies entirely on OODM as in the conceptual data layer. The query system was designed relying on the concept 'find and do'. The query step starts with 'Find', meaning that data objects can be retrieved using the object type followed by the field names beginning with double colon (::); as in the names, _id, inchi, etc. Search keywords can be a full keyword in the double quote symbol ('keyword') or a part of the keyword in-between the slash signs (/part of keyword/). Following with the 'Do' step, related objects can be reached by using object behaviors beginning with dot (.), such as transcribe, translate, catalyze.

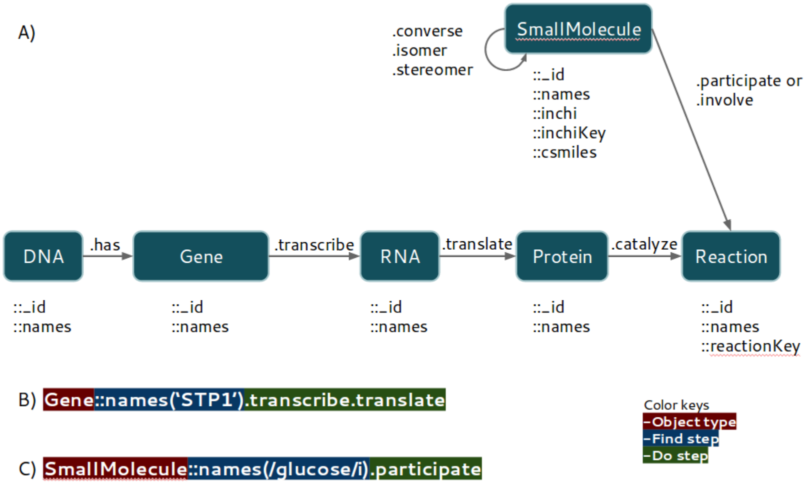


Figure 3.6 *The behavioral query system using Dactyls. (A) The object query structure with the 'Find' step (finding functions beginning with ::) and 'Do' step (beginning with .) functions. The concepts of the system are 'Find' for the interested object and 'Do' for its functions. (B) An example in searching for some proteins that are related to the gene named 'STP1'. (C) An example in searching for the reactions that catalyze glucose.*

Results are returned in an array. The full description of query steps and examples are denoted in Figure 3.6.

3.2.3 Database activities

Connection to the MongoDB DBMS is managed by the ORM MongoModel library, which fully supports the ACID properties for general CRUD only for each object operation.

Create To ensure the integrity of the data, validity rules including uniqueness, required fields and indexes can be applied to the object attributes. Only objects that can pass the validity check and only relation objects with unbroken links can be inserted into the database.

Read The query system relies on the behavioral query system as denoted in the external data layer.

Update In order to update the data, the target object needs to be queried first. All changes have to pass the validity check, same as when inserted before being committed to the database

Delete Objects can be deleted by using the delete command.

4 Visualization platform for multi-omic data on metabolic pathway

Visualization of large-scale biological pathways such as the metabolic pathways is important for the interpretation of multilevel omics data in a pathway context. It is, however, challenging to develop a software tool that is sufficient and flexible to handle multiple metabolic maps and enables the integration of data from different levels. Several metabolic pathway visualization platforms such as the KEGG Atlas (Okuda et al. 2008), Pathway Projector (Kono et al. 2009), iPath (Yamada et al. 2011), BioCyc (Latendresse and Karp 2011) and GLAMM (Bates et al. 2011) have been developed, and they offer zoomable user interface (ZUI) to explore metabolic pathways and allow the users to overlay data on the pathway for further visualization of a specific metabolic pathway. This visualization system was made to facilitate implementation and installation for private applications or public web-services. Besides, data visualization features, Ondine also provides gene set enrichment analysis to determine which pathways are statistically significantly different in terms of gene expression or protein level between two conditions. The Pi-value scoring method was implemented for selecting significant genes and hypergeometric testing was used for gene set enrichment analysis as described below.

4.1 Pi-value

Xiao et al. 2012 introduced a newly developed gene significance score especially for differential gene expression selection for gene set enrichment analysis in order to overcome two problems, 'small fold change, small variance' (SFSV) and 'large fold change, large variance' (LFLV), that can possibly occur when using p-value based selection.

$$\pi_i = \phi_i.(-log_{10}p_i) \tag{4.1}$$

Equation 4.1 illustrates a posterior fusion scheme to combine p-value and fold-change into a new scoring system π, where ϕ_i and p_i are absolute log fold change and p-value respectively. π-value is non negative value, in which genes with larger values are more significant.

4.2 Gene set enrichment analysis

Pathway-gene enrichment analysis is generally estimated using hypergeometric distribution as shown in Equation 4.2. However, due to the integer range limitation of the programming language, the combination was estimated using the Gamma function as shown in Equation 4.3. The function was implemented in the Sylfy library.

$$p_i = \frac{\binom{K}{k_i}\binom{N-K}{n_i-k_i}}{\binom{N}{n_i}} \tag{4.2}$$

where

$$\begin{aligned}
p_i &= \text{p-value of pathway i} \\
N &= \text{Total number of genes} \\
K &= \text{Number of selected genes} \\
n_i &= \text{Number of genes in pathway i} \\
k_i &= \text{Number of selected genes in pathway i} \\
ln\Gamma(n) &= ln((n-1)!) \\
ln(\begin{pmatrix} a \\ b \end{pmatrix}) &= ln\Gamma(a+1) - ln\Gamma(a-b+1) - ln\Gamma(b+1) \qquad (4.3)
\end{aligned}$$

4.3 Ondine visualization engine development

The engine was developed using JavaScript language to render maps in SVG format from map coordinates to the web browser on the client site where the data can be overlaid. Visualization by the Ondine engine is suitable for most of the common web browsers that support SVG and JavaScript such as Firefox, Safari, Chrome, Opera and Internet Explorer without any additional browser add-on requirement. The considered metabolic map will be attached with a HTML div tag, which makes it easy to incorporate into any web design template and JavaScript libraries such as JQueryUI and Prototype UI, should the developer wish to build a more complex web page interface with this engine.

Recently, the KGML format has been developed to manage the coordinates of attributes in the metabolic maps of the KEGG database (Kanehisa et al. 2012). This is the only well-defined data interchange format that provides an easy way to integrate data with the automatic map drawing protocol. It is therefore widely used for many applications. However, KGML is derived from the XML format, which is rarely used in JavaScript. To overcome this problem, map information from KGML has to be converted into the newly developed MJSON format. MJSON relies on JavaScript Object Notation (JSON), which is supported in all JavaScript run-time engines. A map converter is provided by the Sylfy library. The Ondine engine supports any maps generated using layouts derived from the KGML format, which are downloadable from the KEGG database. In the interactive interface architecture, the map components were redrawn as SVG components using the D3js library working on backend in response to the users' actions as shown in Figure 4.1.

4.4 Ondine web service implementation

To make the understanding of multilevel high throughput data effortless and more meaningful, a visualization system is therefore needed. Ondine (Omics navigator for data

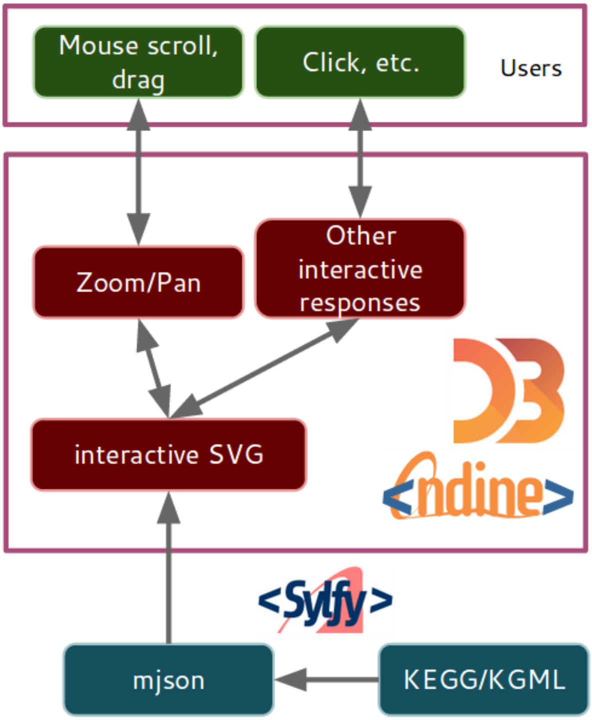
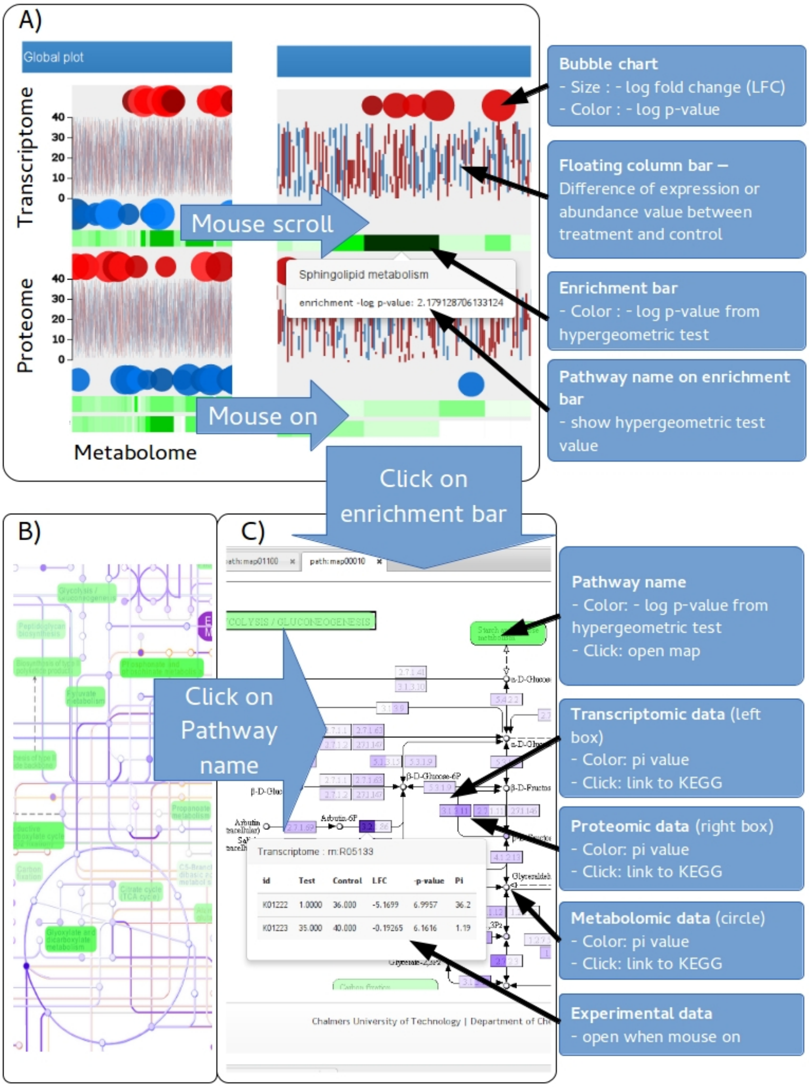


Figure 4.1 *The Ondine engine is a JavaScript library built using the D3js library to render SVG metabolic maps, which provide interactive responses, zoom, pan and popup windows to the users. (Adapted from* ***Paper I****)*

integration and evaluation) was developed to interactively visualize multilevel omics data, which allows users to simultaneously navigate the transcriptomic, proteomic and metabolomic data in the KEGG pathway. Ondine is a web application developed in Ruby on Rails and JavaScript. Ondine's backend was also used to interactively visualize human tissue specific genome scale metabolic models in the Human Metabolic Atlas website. All these software developments are intended to support data expansion and utilization not only for the Human Metabolic Atlas project, but also hopefully for the global research community.

The global plot was developed and adapted from the cloud plots of global metabolic data (Patti et al. 2013) as a compact data illustration that simultaneously represents expression values and statistical analysis of the transcriptome and proteome with gene set enrichment analysis on the KEGG pathway gene set. The plot is comprised of the differentiate bar, circle color and size which represents the differential fold change between the treatment and control, the fold change and significant value respectively. The enrichment bar represents the significant values of the enrichment analysis for each pathway, which can provide an overview of the significant pathways related to the experiment data.

An example plot is shown in Figure 4.2 (A). The global metabolic map is opened for giving an overview of the relationship between the experimental data and the metabolic pathway by default. Others can be opened by clicking on the enrichment bar or the pathway name from another map. Ondine provides the interfaces that allow the users to interactively explore the map and its component details and to overlay data by using simple mouse controls. By mouse scrolling, the user can smoothly zoom in and out without loss of resolution from the original picture. Details of the map components including the id, name, link and graphs of the overlaid data can be shown in a balloon popup by performing a left mouse click on the component as shown in Figure 4.2 (B).


A)
Global plot
Transcriptome
Proteome
Metabolome
40
30
20
10
0
Mouse scroll
Mouse on
Sphingolipid metabolism
enrichment -log p-value: 2.179128706133124
Bubble chart
- Size : - log fold change (LFC)
- Color : - log p-value
Floating column bar –
Difference of expression or abundance value between treatment and control
Enrichment bar
- Color : - log p-value from hypergeometric test
Pathway name on enrichment bar
- show hypergeometric test value
Click on enrichment bar
B)
C)
path:map01100
path:map00010
Click on Pathway name
Pathway name
- Color: - log p-value from hypergeometric test
- Click: open map
Transcriptomic data (left box)
- Color: pi value
- Click: link to KEGG
Proteomic data (right box)
- Color: pi value
- Click: link to KEGG
Metabolomic data (circle)
- Color: pi value
- Click: link to KEGG
Experimental data
- open when mouse on
Chalmers University of Technology | Department of Che

Transcriptome : rn:R05133

id	Test	Control	LFC	-p-value	Pi
K01222	1.0000	36.000	-5.1699	6.9957	36.2
K01223	35.000	40.000	-0.19265	6.1616	1.19

Figure 4.2 *Ondine web service provides a global plot view and data overlaid metabolic pathway maps for illustrating multi-omic data that focus on the metabolic viewpoint. A) The global plot represents the difference in expression values between the treatment and control, the p-value of differential expression, the log-fold change and pathway enrichment analysis. B) The global metabolic map is opened by default to present an overview of the data. C) Sub-maps, which can be opened by clicking on the pathway name, represent the information of specific pathways. (From **Paper I**)*

5 Human Metabolic Atlas website

In order to efficiently manage and utilize GEM and related information, the Human Metabolic Atlas (HMA) website was built as an online resource to provide comprehensive human metabolic information for supporting further specific analysis or modeling as well as to communicate with the wider research community. The website was developed mainly using the Ruby on Rails platform. The latest version of the HMA is comprised of 3 parts: the repositories, the Hreed database and an Atlas as shown in Figure 5.1.

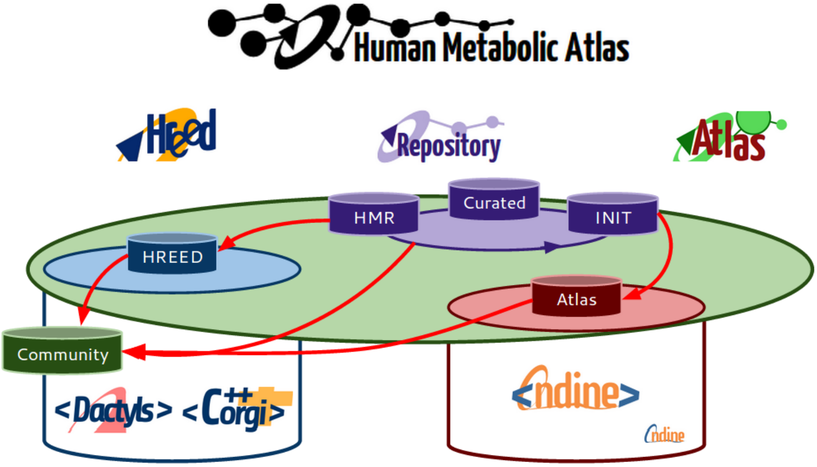
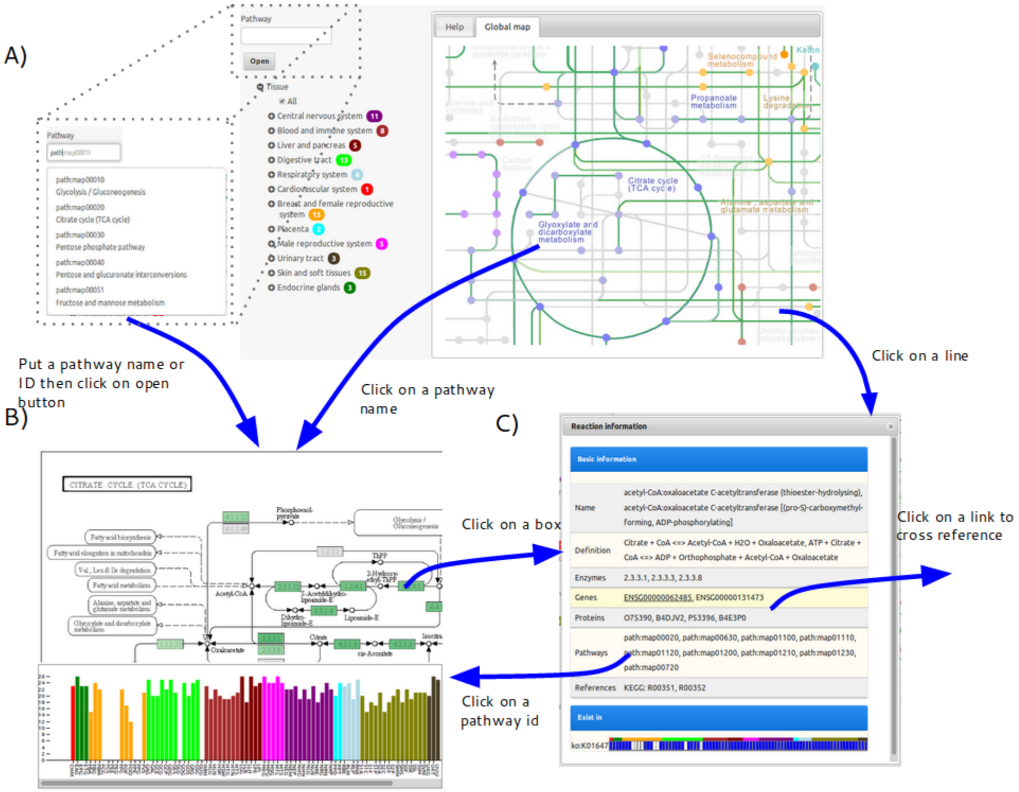


Figure 5.1 *Three applications of HMA website is aimed to provide human metabolism to research community. Users can accesses data directly from Repository by downloading models in SBML or using map viewer and query system.*

The repositories provide 99 human tissue specific models including normal, cancer and curated models and 3 human-related microbial models that are available for downloading in SBML format. With the available repositories, the HMA can be considered as a comprehensive web resource for 1) providing draft GEMs for both normal and cancer cell types generated by the automatic algorithm INIT; 2) providing simulation ready and functional GEMs, which can serve as a prediction model and scaffold for personalizing genome scale metabolic models, which can both significantly contribute to the understanding of diseases prior to finding the therapeutics.

To support data exchange and the expansion of human metabolism knowledge, the Hreed database, considered as an initial tool set, was attentively developed using the Corgi and Dactyls database API. Hreed was initially automatically propagated from HMR using Corgi API to ensure accuracy and integrity of data. To provide a graphical interface to the Hreed database, the web-based data query system was developed as shown in Figure 5.3. Data objects, including genes, transcripts, proteins, reactions and small molecules, can be queried by using simple keywords such as names, id, cross references, InChI and InChIKey. Besides using a string of keywords, regular expression and wild cards such as '.', '*' can also be used for creating a complex search term. Related objects can be further queried by specifying the relation types. Query results are presented on the web as a table view by default. Cross reference links are also provided in the table view allowing the user to further explore the details either from the database itself or from external databases. The results can be downloaded as a text file in a table, XML and JSON format, which is more convenient for using the data for further computational analysis.

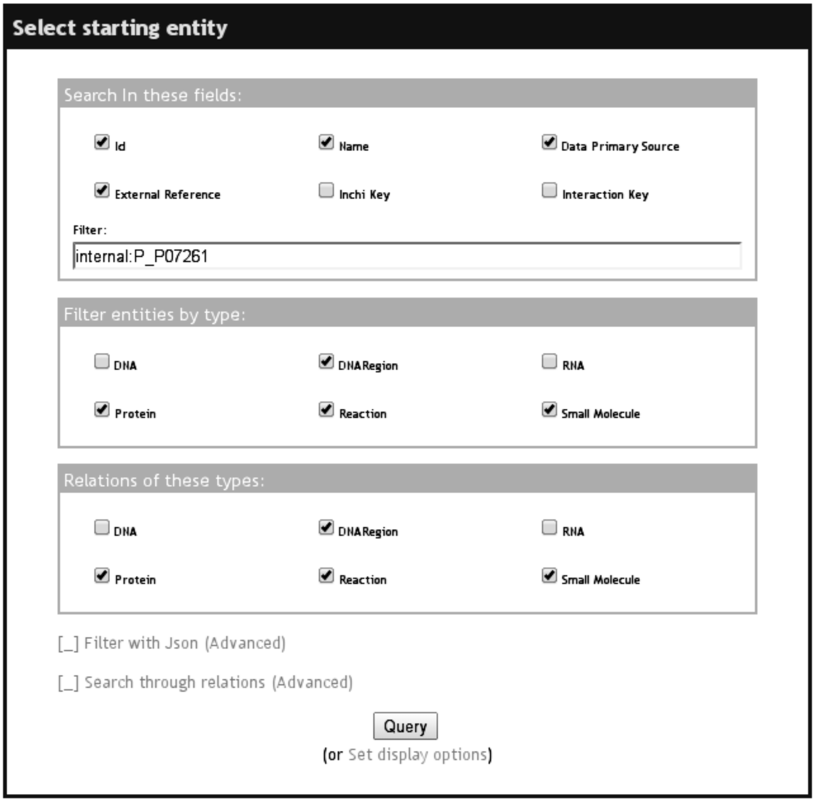
The Atlas was implemented using the Ondine engine to provide a comparative view of several GEMs on a metabolic map. This map will be useful for observing the relationship and differentiation for each cell type in a graphical way that is easier to understand. Gene lists from each INIT GEM were mapped to the KEGG gene list for each pathway to generate the map coordinate data in MJSON format, which can be interactively rendered on the web browser by the Ondine engine (**Paper I**). The map can be opened by typing into the auto-filled input box or by clicking on the map name on other maps. Overlaid tissue information can be chosen from a tissue filter tree on the left hand side of the map. In the viewer, the summary of the gene number for each tissue is shown in a bar chart under the map. Further information can be presented in a popup, which will be shown after clicking. The full information is illustrated in Figure 5.2.



A)
Pathway
Open
Tissue
☑ All
Central nervous system 11
Blood and immune system 8
Liver and pancreas 5
Digestive tract 13
Respiratory system 4
Cardiovascular system 1
Breast and female reproductive system 13
Placenta 2
Male reproductive system 5
Urinary tract 3
Skin and soft tissues 15
Endocrine glands 3
Pathway
path:map00010
path:map00010
Glycolysis / Gluconeogenesis
path:map00020
Citrate cycle (TCA cycle)
path:map00030
Pentose phosphate pathway
path:map00040
Pentose and glucuronate interconversions
path:map00051
Fructose and mannose metabolism
Help
Global map
Selenocompound metabolism
Propanoate metabolism
Lysine degradation
Citrate cycle (TCA cycle)
Glyoxylate and dicarboxylate metabolism
Alanine, aspartate and glutamate metabolism
Put a pathway name or ID then click on open button
Click on a pathway name
Click on a line
B)
CITRATE CYCLE (TCA CYCLE)
Glycolysis / Gluconeogenesis
Fatty acid biosynthesis
Fatty acid elongation in mitochondria
Val, Leu & Ile degradation
Fatty acid metabolism
Alanine, aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
Phosphoenol-pyruvate
Acetyl-CoA
Oxaloacetate
Citrate
cis-Aconitate
Isocitrate
TPP
2-Hydroxy-ethyl-TPP
S-Acetyldihydro-lipoamide-E
Dihydro-lipoamide-E
Lipoamide-E
C)
Click on a box
Click on a link to cross reference
Click on a pathway id
Reaction information
Basic information
Exist in
ko:K01647

Name	acetyl-CoA:oxaloacetate C-acetyltransferase (thioester-hydrolysing), acetyl-CoA:oxaloacetate C-acetyltransferase [(pro-S)-carboxymethyl-forming, ADP-phosphorylating]
Definition	Citrate + CoA <=> Acetyl-CoA + H2O + Oxaloacetate, ATP + Citrate + CoA <=> ADP + Orthophosphate + Acetyl-CoA + Oxaloacetate
Enzymes	2.3.3.1, 2.3.3.3, 2.3.3.8
Genes	ENSG00000062485, ENSG00000131473
Proteins	O75390, B4DJV2, P53396, B4E3P0
Pathways	path:map00020, path:map00630, path:map01100, path:map01110, path:map01120, path:map01200, path:map01210, path:map01230, path:map00720
References	KEGG: R00351, R00352

Figure 5.2 *Summary of Atlas functions. (A) The main panel of Atlas on the left hand side is comprised of pathway input box for selecting specific metabolic pathway map to be opened and an tissue option tree categorized by system of for selecting cell type information to be overlaid on the map. Global metabolic pathway map is opened by default. (B) Sub metabolic map with data overlaid and bar plot representing the number of genes that represents in the pathway map for each cell type, which can be opened by control panel, clicking on pathway name in every map and clicking on pathway id in information window. (C) The information window, popped up when clicking on a map component, represents information of the reaction from KEGG and provides link to external data sources for further information.*



Select starting entity

Search In these fields:

☑ Id ☑ Name ☑ Data Primary Source

☑ External Reference ☐ Inchi Key ☐ Interaction Key

Filter: internal:P_P07261

Filter entities by type:

☐ DNA ☑ DNARegion ☐ RNA

☑ Protein ☑ Reaction ☑ Small Molecule

Relations of these types:

☐ DNA ☑ DNARegion ☐ RNA

☑ Protein ☑ Reaction ☑ Small Molecule

[_] Filter with Json (Advanced)

[_] Search through relations (Advanced)

Figure 5.3 *Web-based data query system. Users simply provide a keyword into the filter input box and choose the specific fields to further specify the search criteria. Two or more keywords can be combined together with the '—' (or) and '+' (and) operators. The results can be refined by using the 'filter entities by type' and 'relations of these types' tabs in order to obtain specific data types or their related data objects.*

6 Summaries and perspectives

In order to manage the high complexity of the data used in systems biology, particularly for GEM reconstruction, two specific database API libraries for handling data with two different purposes were presented. Corgi is mainly for programmers, while Dactyls is intended for the end user to connect to the database. They allow and support crucial tasks in areas including integration and analysis of multi-level omics data, modeling of cellular pathways and collecting biological network data. The libraries provide essential classes and services for communication among the layers. The basic properties of a database system, ACID, were considered by providing specific functions and control processes in the library to ensure that the database transactions and the data inside are consistent, reliable and non corrupted.

As a major part of the database system development, the data schema was designed based on NO-SQL data models using a three data layers concept. The concept divided the data structure into 3 layers consisting of the physical, conceptual and external layers, to allow for independent and effective implementation or changes at each data layer. The physical layer is the document-oriented data model, which was implemented in the underline database management system, MongoDB. The conceptual layer is implemented in the database API libraries, both Corgi and Dactyls. An object-oriented concept was adopted for the design of the conceptual data structure since it represents real world information as an object with related attributes and a variety of relationships. It can make the manipulation of data that are regarded as objects and their related information easier, more straight forward and relatively faster. In addition, the concept is applicable for capturing and reflecting biological information that are apparently heterogeneous and sophisticated Okayama et al. 1998. The major design of the conceptual data structure that characterizes data in systems biology was adapted from BioPAX ontology (Demir et al. 2010). To provide an effortless data query system, a behavioral-based data query system was implemented in Dactyls as an external layer of the database system. A query key word can be constructed by using simple molecular biology knowledge regardless of the conceptual data structure. By realizing the usages of different standard formats, the parser classes were included in the library. These classes support the standard formats that are generally used in most biological databases to accommodate the integration of data from different sources into the database and to enhance the extensibility of the data structure.

The API libraries show an extensive attempt to serve and solve complex data handling and integration in systems biology by following and using different standards and technologies. It provides users with the ability to personalize the views of data through additional applications and ensures the integrity, consistency and reliability of data in the database. Besides the general features of database management, the database system was designed to be extensible and easy integrated with the upcoming technology in the database management field. With the current situation in the informatics era, where an enormous amount of information is being generated and becoming publicly available in the internet network, the way to manage and analyze data are moving forward towards relying more on the data itself. The database system can be easily extended, for which modern data analysis approaches such as the data centric analysis (Chodorow 2013; Quintero et al. 2013) and context-aware data query system (Feng et al. 2004) can potentially be applied to.

Besides data management work, in order to deliver information from the database to the end users (i.e. biologists), a visualization system is needed. Visualization of omic data on metabolic pathways is important to capture an overview of the large metabolic system. The Ondine engine is a useful tool for the visualization of multi-omic data including transcriptome, proteome and metabolome simultaneously on biological network maps and it enables easy data integration and has an interactive ZUI feature. Metabolic maps are rendered as a SVG image, using only the coordinates of shapes and binding information, which are easily generated from most of the metabolic map file formats. In this particular work, map information files are generated from the KEGG KGML and KGML+ files. Other pathway maps from Wikipathway (Kelder et al. 2012), Reactome (Croft et al. 2014) and PID (Schaefer et al. 2009) databases will be populated and would be available for navigation in the future.

In recent years, human tissue-specific genome-scale metabolic (GEM) modelling has provided many new information about human metabolism with the integration of genomic, transcriptomic, proteomic, metabolomic and biochemical reaction information. To efficiently manage and utilize these highly complex data existing in the genome scale metabolic models remain challenging. With the newly developed database and visualization system, the Human Metabolic Atlas was built to serve as an online resource to support data exchange among the human metabolic research community. Repositories provide GEM in the widely supported SBML format, while Hreed provides well-annotated and standardized human reaction data with a user interface. However, the HMA is still an ongoing development aiming to provide more features such as GEM reconstruction support in the future.

Acknowledgements

First of all I would like to express my special appreciation and thanks to my supervisor Jens Nielsen for giving me a great opportunity to pursue a PhD study under your supervision. Your advices on both my research as well as on my career have been invaluable.

Thanks to my co-supervisor Intawat Nookaew for a lot of ideas, discussion and a good taking care for all whole four years here. Your guidances helped me in all the time of research and writing of this thesis.

I would especially like to thank all people who have contributed to this work. Without them, this work might not be done. Kwanjeera Wanichthanarak has been a great collaborator. Rasmus Ågren, Avlant Nilsson, Adil Mardinoglu and Sergio Bordel were greatly contributing in this work. Thanks go to all computational people, Tobias Österlund, Leif Väremo, Fredrik Karlsson, Francesco Gatto, Saeed Shoaie, Shaghayegh (Shaq) Hosseini and Luis Caspeta-Guadarrama for great scientific discussion. A special thank goes to Tawan Thammasunthorn for the thesis proofreading.

Furthermore I would also like to acknowledge with much appreciation to the Thai community in Sysbio group, Sakda Khoomrung, Kanokarn Kocharin, Pramote Chumnanpuen, Suwanee Jansa-Ard and Kanthida Kusonmano for all good advices and friendships. Thanks go to Christoph Knuf, Bouke de Jong, Nina Johansson, Martin Markström and Verena Siewer for spending happy time together. I would like to thank all my office mates, Lifang Liu, Juan Octavio Valle Rodriguez, Ana Joyce Muñoz Arellano and Rahul Kumar. Special thanks go to Erica Dahlin and Helena Janveden for administrative supports.

In addition, I would like to express my special thanks to Chinae Thammarongtham, Pravech Ajawatanawong, Aj. Charassri Chinnawong and Aj. Duang-deun Meksuriyen for giving support and encouragement to pursue my study. Thanks go to Wanwipa Vongsangnak for introducing me to Jens.

I would like to give special thanks to all Thai students in Göteborg especially Nutcha Teraphongphom, Phasin Ratanalert and Piriya Tangnavarad for sharing the lively time together. Many special thanks are also given to my friends for encouraging me during my study.

Last but not least, a special thanks to my family. Words cannot express how grateful I am to my parents and my siblings for their unconditional support. At the end I would like express appreciation to my loved one who supports me and always stays with me.

References

Agren, R., S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, and J. Nielsen (2012). Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT. *PLoS Computational Biology* **8**.5. Ed. by C. D. Maranas, e1002518. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1002518. URL: http://dx.plos.org/10.1371/journal.pcbi.1002518.

Bachrach, S. M. (Jan. 2012). InChI: a user's perspective. *Journal of cheminformatics* **4**.1, 34. ISSN: 1758-2946. DOI: 10.1186/1758-2946-4-34. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3537656%5C&tool=pmcentrez%5C&rendertype=abstract.

Barrett, T. et al. (Jan. 2011). NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic acids research* **39**.Database issue, D1005–10. ISSN: 1362-4962. DOI: 10.1093/nar/gkq1184. URL: http://nar.oxfordjournals.org/content/39/suppl%5C_1/D1005.

Barry, D. (1996). *The object database handbook: how to select, implement, and use object-oriented databases.* New York, {NY}, {USA}: John Wiley &amp; Sons, Inc. ISBN: 0-471-14718-4. URL: http://dl.acm.org/citation.cfm?id=235131.

Bates, J. T., D. Chivian, and A. P. Arkin (July 2011). GLAMM: Genome-Linked Application for Metabolic Maps. *Nucleic acids research* **39**.Web Server issue, W400–5. ISSN: 1362-4962. DOI: 10.1093/nar/gkr433. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125797%5C&tool=pmcentrez%5C&rendertype=abstract.

Benson, D. a., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers (Jan. 2013). GenBank. *Nucleic acids research* **41**.Database issue, D36–42. ISSN: 1362-4962. DOI: 10.1093/nar/gks1195. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531190%5C&tool=pmcentrez%5C&rendertype=abstract.

Bertino, E. and L. Martino (1993). *Object-oriented database systems: concepts and architectures.* Addison-Wesley Pub. Co. ISBN: 9780201624397. URL: http://dl.acm.org/citation.cfm?id=562610.

Birney, E. and M. Clamp (Mar. 2004). Biological database design and implementation. *Briefings in bioinformatics* **5**.1, 31–8. ISSN: 1467-5463. URL: http://www.ncbi.nlm.nih.gov/pubmed/15153304.

Bolton, E., Y. Wang, P. Thiessen, and S. Bryant (2008). PubChem: integrated platform of small molecules and biological activities. *Annual reports in ...* **4**. URL: http://oldwww.acscomp.org/Publications/ARCC/volume4/chapter12.html%20http://www.sciencedirect.com/science/article/pii/S1574140008000121.

Cary, M., G. Bader, and C. Sander (Mar. 2005). Pathway information for systems biology. *FEBS letters* **579**.8, 1815–1820. ISSN: 0014-5793. DOI: 10.1016/j.febslet.2005.02.005. URL: http://www.sciencedirect.com/science/article/pii/S0014579305001705.

Chodorow, K. (2013). *MongoDB: the definitive guide.* URL: http://shop.oreilly.com/product/0636920001096.do%20http://books.google.com/books?hl=en%5C&lr=

%5C&id=pAbSHFi4WSAC%5C&oi=fnd%5C&pg=PA5%5C&dq=MongoDB:+The+Definitive+Guide%5C&ots=Dt8eNzp7y7%5C&sig=QTn-NgQ1ydqFKdLZrQ8ZYMKuoIE.

Croft, D. et al. (Jan. 2014). The Reactome pathway knowledgebase. *Nucleic acids research* **42**.1, D472–7. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1102. URL: http://www.ncbi.nlm.nih.gov/pubmed/24243840.

Davidson, S. (1995). Challenges in integrating biological data sources. *... of Computational Biology* **2**.4, 557–572. URL: http://online.liebertpub.com/doi/abs/10.1089/cmb.1995.2.557.

Demir, E., M. Cary, S. Paley, and K. Fukuda (Sept. 2010). The BioPAX community standard for pathway data sharing. *Nature ...* **28**.9, 935–942. ISSN: 1546-1696. DOI: 10.1038/nbt.1666. URL: http://www.ncbi.nlm.nih.gov/pubmed/20829833%20http://www.nature.com/nbt/journal/v28/n9/abs/nbt.1666.html.

Duarte, N. C., S. a. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson (Feb. 2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America* **104**.6, 1777–82. ISSN: 0027-8424. DOI: 10.1073/pnas.0610772104. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1794290%5C&tool=pmcentrez%5C&rendertype=abstract.

Ecma, S. (June 2009). *ECMA-262 ECMAScript Language Specification.* 5th ed. Ecma International. URL: http://www.ecma-international.org/ecma-262/5.1/%20http://scholar.google.com/scholar?hl=en%5C&btnG=Search%5C&q=intitle:ECMA-262:+ECMAScript+Language+Specification%5C#0%20http://scholar.google.com/scholar?hl=en%5C&btnG=Search%5C&q=intitle:ECMA-262+ECMAScript+Language+Specification%5C#0.

Feng, L., P. Apers, and W. Jonker (2004). Towards context-aware data management for ambient intelligence. *Database and Expert Systems ...* 422–431. URL: http://link.springer.com/chapter/10.1007/978-3-540-30075-5%5C_41.

Grethe, G., J. M. Goodman, and C. H. Allen (Jan. 2013). International chemical identifier for reactions (RInChI). *Journal of cheminformatics* **5**.1, 45. ISSN: 1758-2946. DOI: 10.1186/1758-2946-5-45. URL: http://www.ncbi.nlm.nih.gov/pubmed/24152584.

Hastings, J. et al. (Jan. 2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research* **41**.Database issue, D456–63. ISSN: 1362-4962. DOI: 10.1093/nar/gks1146. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531142%5C&tool=pmcentrez%5C&rendertype=abstract.

Heller, S., A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev (Jan. 2013). InChI - the worldwide chemical structure identifier standard. *Journal of cheminformatics* **5**.1, 7. ISSN: 1758-2946. DOI: 10.1186/1758-2946-5-7. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3599061%5C&tool=pmcentrez%5C&rendertype=abstract.

Hoffer, J. A., V. Ramesh, and H. Topi (2011). *Modern database management.* 10th ed. Prentice Hall. ISBN: 9780136088394. URL: http://www.pearsonhighered.com/bookseller/product/Modern-Database-Management-10E/9780136088394.page.

Ideker, T., T. Galitski, and L. Hood (2001). A new approach to decoding life: systems biology. *Annual review of genomics and ...* URL: http://www.annualreviews.org/doi/abs/10.1146/annurev.genom.2.1.343.

Jerby, L., T. Shlomi, and E. Ruppin (Sept. 2010). Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Molecular systems biology* **6**.401, 401. ISSN: 1744-4292. DOI: 10.1038/msb.2010.56. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2964116%5C&tool=pmcentrez%5C&rendertype=abstract.

Juty, N., N. Le Novère, and C. Laibe (Jan. 2012). Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic acids research* **40**.Database issue, D580–6. ISSN: 1362-4962. DOI: 10.1093/nar/gkr1097. URL: http://nar.oxfordjournals.org/cgi/content/long/40/D1/D580.

Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe (Jan. 2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40**.Database issue, D109–14. ISSN: 1362-4962. DOI: 10.1093/nar/gkr988. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245020%5C&tool=pmcentrez%5C&rendertype=abstract.

Kelder, T., M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico (Jan. 2012). WikiPathways: building research communities on biological pathways. *Nucleic acids research* **40**.Database issue, D1301–7. ISSN: 1362-4962. DOI: 10.1093/nar/gkr1074. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245032%5C&tool=pmcentrez%5C&rendertype=abstract.

Kitano, H. (Nov. 2002). Computational systems biology. *Nature* **420**.6912, 206–10. ISSN: 0028-0836. DOI: 10.1038/nature01254. URL: http://www.ncbi.nlm.nih.gov/pubmed/17052114.

Kono, N., K. Arakawa, R. Ogawa, N. Kido, K. Oshita, K. Ikegami, S. Tamaki, and M. Tomita (Jan. 2009). Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. *PloS one* **4**.11, e7710. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0007710. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2770834%5C&tool=pmcentrez%5C&rendertype=abstract.

Kumar, A., P. F. Suthers, and C. D. Maranas (Jan. 2012). MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC bioinformatics* **13**.1, 6. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-6. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3277463%5C&tool=pmcentrez%5C&rendertype=abstract.

Latendresse, M. and P. D. Karp (Jan. 2011). Web-based metabolic network visualization with a zooming user interface. *BMC bioinformatics* **12**.1, 176. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-176. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3113945%5C&tool=pmcentrez%5C&rendertype=abstract.

Ma, H., A. Sorokin, A. Mazein, A. Selkov, E. Selkov, O. Demin, and I. Goryanin (Jan. 2007). The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular systems biology* **3**.135, 135. ISSN: 1744-4292. DOI: 10.1038/msb4100177.

URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2013923%5C&tool=pmcentrez%5C&rendertype=abstract.

Mardinoglu, A. et al. (Jan. 2013). Integration of clinical data with a genome-scale metabolic model of the human adipocyte. en. *Molecular systems biology* **9**.1, 649. ISSN: 1744-4292. DOI: 10.1038/msb.2013.5. URL: http://msb.embopress.org/content/9/1/649.abstract.

Millard, B. L., M. Niepel, M. P. Menden, J. L. Muhlich, and P. K. Sorger (June 2011). Adaptive informatics for multifactorial and high-content biological data. *Nature methods* **8**.6, 487–93. ISSN: 1548-7105. DOI: 10.1038/nmeth.1600. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3105758%5C&tool=pmcentrez%5C&rendertype=abstract.

Nookaew, I., M. Papini, N. Pornputtpong, G. Scalcinati, L. Fagerberg, M. Uhlén, and J. Nielsen (Sept. 2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic acids research* **40**.20, 10084–10097. ISSN: 1362-4962. DOI: 10.1093/nar/gks804. URL: http://nar.oxfordjournals.org/content/40/20/10084.

Object Management Group (Aug. 2011). *Documents associated with Unified Modeling Language (UML), v2.4.1.* Object Management Group Inc.

Okayama, T., T. Tamura, T. Gojobori, Y. Tateno, K. Ikeo, S. Miyazaki, K. Fukami-Kobayashi, and H. Sugawara (July 1998). Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library. *Bioinformatics* **14**.6, 472–478. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/14.6.472. URL: http://bioinformatics.oxfordjournals.org/cgi/content/long/14/6/472.

Okuda, S., T. Yamada, M. Hamajima, M. Itoh, T. Katayama, P. Bork, S. Goto, and M. Kanehisa (July 2008). KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic acids research* **36**.Web Server issue, W423–6. ISSN: 1362-4962. DOI: 10.1093/nar/gkn282. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447737%5C&tool=pmcentrez%5C&rendertype=abstract.

Ozsoyoglu, Z., G. Ozsoyoglu, and J. Nadeau (Aug. 2006). Genomic pathways database and biological data management. *Animal genetics* **37 Suppl 1**, 41–47. ISSN: 0268-9146. DOI: 10.1111/j.1365-2052.2006.01477.x. URL: http://www.ncbi.nlm.nih.gov/pubmed/16887001%20http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2052.2006.01477.x/full.

Patti, G. J., R. Tautenhahn, D. Rinehart, K. Cho, L. P. Shriver, M. Manchester, I. Nikolskiy, C. H. Johnson, N. G. Mahieu, and G. Siuzdak (Jan. 2013). A view from above: cloud plots to visualize global metabolomic data. *Analytical chemistry* **85**.2, 798–804. ISSN: 1520-6882. DOI: 10.1021/ac3029745. URL: http://www.ncbi.nlm.nih.gov/pubmed/23206250.

Pletnev, I., A. Erin, A. McNaught, K. Blinov, D. Tchekhovskoi, and S. Heller (Jan. 2012). InChIKey collision resistance: an experimental testing. *Journal of cheminformatics* **4**.1, 39. ISSN: 1758-2946. DOI: 10.1186/1758-2946-4-39. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3558395%5C&tool=pmcentrez%5C&rendertype=abstract.

Quintero, C., K. Tran, and A. a. Szewczak (Aug. 2013). High-throughput quality control of DMSO acoustic dispensing using photometric dye methods. *Journal of laboratory automation* **18**.4, 296–305. ISSN: 2211-0682. DOI: 10.1177/2211068213486787. URL: http://www.ncbi.nlm.nih.gov/pubmed/23629143.

Schaefer, C. F., K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow (Jan. 2009). PID: the Pathway Interaction Database. *Nucleic acids research* **37**.Database issue, D674–9. ISSN: 1362-4962. DOI: 10.1093/nar/gkn653. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686461%5C&tool=pmcentrez%5C&rendertype=abstract.

Steel (jr.), T. B. (1975). *Interim Report: ANSI/X3/SPARC Study Group on Data Base Management Systems.* American National Standards Institute. URL: http://scholar.google.com/scholar?hl=en%5C&btnG=Search%5C&q=intitle:Interim+Report+ANSI/X3/SPARC+Study+Group+on+Data+Base+Management+Systems%5C#2.

Sud, M. et al. (Jan. 2007). LMSD: LIPID MAPS structure database. *Nucleic acids research* **35**.Database issue, D527–32. ISSN: 1362-4962. DOI: 10.1093/nar/gkl838. URL: http://nar.oxfordjournals.org/content/35/suppl%5C_1/D527.abstract.

Tao, Y., Y. Liu, C. Friedman, and Y. Lussier (2004). Information visualization techniques in bioinformatics during the postgenomic era. *Drug Discovery Today: BIOSILICO* **2**.6, 237–245. DOI: 10.1016/S1741-8364(04)02423-0.Information. URL: http://www.sciencedirect.com/science/article/pii/S1741836404024230.

The UniProt Consortium (Jan. 2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research* **41**.Database issue, D43–7. ISSN: 1362-4962. DOI: 10.1093/nar/gks1068. URL: http://nar.oxfordjournals.org/content/41/D1/D43.

Wang, Y., J. a. Eddy, and N. D. Price (Jan. 2012). Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC systems biology* **6**, 153. ISSN: 1752-0509. DOI: 10.1186/1752-0509-6-153. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3576361%5C&tool=pmcentrez%5C&rendertype=abstract.

Whetzel, P. and N. Noy (July 2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic acids ...* **39**.Web Server issue, W541–W545. ISSN: 0305-1048. DOI: 10.1093/nar/gkr469. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125807/%20http://nar.oxfordjournals.org/content/39/suppl%5C_2/W541.short.

Williams, A. J. (Jan. 2012). InChI: connecting and navigating chemistry. *Journal of cheminformatics* **4**.1, 33. ISSN: 1758-2946. DOI: 10.1186/1758-2946-4-33. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3537679%5C&tool=pmcentrez%5C&rendertype=abstract.

Wishart, D. S., C. Knox, et al. (Jan. 2009). HMDB: a knowledgebase for the human metabolome. *Nucleic acids research* **37**.Database issue, D603–10. ISSN: 1362-4962. DOI: 10.1093/nar/gkn810. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686599%5C&tool=pmcentrez%5C&rendertype=abstract.

Wishart, D. S., D. Tzur, et al. (Jan. 2007). HMDB: the Human Metabolome Database. *Nucleic acids research* **35**.Database issue, D521–6. ISSN: 1362-4962. DOI: 10.1093/

nar/gkl923. URL: http://nar.oxfordjournals.org/content/35/suppl%5C_1/D521.abstract.

Xiao, Y., T.-H. Hsiao, U. Suresh, H.-I. H. Chen, X. Wu, S. E. Wolf, and Y. Chen (Feb. 2012). A Novel Significance Score for Gene Selection and Ranking. *Bioinformatics (Oxford, England)*, btr671–. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr671. URL: http://bioinformatics.oxfordjournals.org/cgi/content/long/btr671v1.

Yamada, T., I. Letunic, S. Okuda, M. Kanehisa, and P. Bork (July 2011). iPath2.0: interactive pathway explorer. *Nucleic acids research* **39**.Web Server issue, W412–5. ISSN: 1362-4962. DOI: 10.1093/nar/gkr313. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125749%5C&tool=pmcentrez%5C&rendertype=abstract.

Zhao, L. and S. A. Roberts (Feb. 1988). An Object-Oriented Data Model for Database Modelling, Implementation and Access. *The Computer Journal* **31**.2, 116–124. ISSN: 0010-4620. DOI: 10.1093/comjnl/31.2.116. URL: http://comjnl.oxfordjournals.org/content/31/2/116.abstract.

Part I
Appended Papers I-IV

Paper I

Ondine: A web application for multilevel omics data integration and visualization

Ondine: A web application for multilevel omics data integration and visualization

Natapol Pornputtapong[1], Jens Nielsen[1] and Intawat Nookaew [1,2,*]

[1]Department of Chemical and Biological Engineering, Chalmers University of Technology, Sweden

[2] Current address : Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

*corresponding author

email: intawat@chalmers.se

Fax: +46 (31) 772 38 01

Tel: +46 (31) 772 38 54

NP's email: natapol@chalmers.se
JN's email: nielsenj@chalmers.se
IN's email: intawat@chalmers.se

Abstract

Background

With the capability of high throughput technologies, probing cellular process can be routinely accomplished through transcriptomic, proteomic and metabolomic data. There is need a tool for data integration and visualization of the multilevel omics data.

Results

Ondine (Omics navigator for data integration and evaluation) is developed to be a web application for data integration and visualization of multilevel omics data on pathway maps, simultaneously. Ondine provides JavaScript based interactive interface with many useful features enabling users to summarize the uploaded multilevel omics data in statistical global plots with pathway enrichment analysis and to navigate the overlaid omics data on KEGG global maps and specific pathway maps. Users only supply their data then all the provided maps will be automatically rendered and ready to navigate.

Conclusion

Under Ondine environment, systematic data exploration, mining and integration of multilevel omics data can be performed in concert. Ondine web application is freely available online at www.ondine.se

Background

Emerging of high throughput technologies have enabled research scientist to monitor cellular changes in genome-wide fashion as called omics data in routine. This leads to the rapid accumulation and sharing of multilevel omics data in the public repository databases. However, only ~20 % of the publicly available data has been referred in the other research work [1]. This strongly indicates the needs for efficient utilization of the highly complex data in the research community. Analysis and understanding of the multilevel omics data can be accomplished through advance visualization that provides clear, meaningful and integrative environment [2]. This will create “emergent properties” toward the biological interpretation, insight and discovery. Most of existing web applications for metabolic pathway visualization platform have been developed to meet the mention objectives, such as KEGG Atlas [3], Pathway Projector [4], iPath [5], BioCyc [6] and GLAMM [7]. They offer zoomable user interface (ZUI) to explore metabolic network and allow user to integrate individual omics data on the global map.

Herein, we provided a web application for concerted omics data integration and visualization of transcriptomic, proteomic and metabolomic data, as called Ondine (**O**mics **n**avigator for **d**ata **in**tegration and **e**valuation). Ondine is designed to be a simple and lightweight ZUI equipped with many useful interactive features that assist the user to easily navigate and evaluate their multilevel omics data on the provided KEGG maps covered both metabolic and non-metabolic pathway maps.

Implementation

Ondine is developed as a web application on Ruby, Rails platform and using JavaScript with AJAX technology to rapidly illustrate global plot and maps into any web browser. The interactive interface allows users to inter-actively explore the plot, map, component details and overlaid data by using simple mouse controls. By mouse scrolling and clicking, user can smoothly zoom in and out without resolution loss and show detailed data with external links respectively. Ondine is fully compatible with all web browsers without any plug-ins installation. Rendering of the map and plot of all components as well as interactive response is computed on the user computer. The summary of Ondine architecture is summarized in Figure 1. In brief, maps (KGML/KGML+) will be firstly converted to mjson format. The map will be interactive rendered under SVG following the user interactions.

Results

All key interactive features of Ondine are summarized in Figure 2 and the description in details are summarize below

Data Preparation and ID Conversion

With Ondine, users can easily overview any experimental data including expression level of transcripts and proteins or detected level of metabolites on the graphic components through ids mapping between map information and a data file provided by user. The input file, in tab-separated values (TSV) format, is comprised of 5 columns, 'class', 'id', 'test', 'control' and 'pvalue' respectively. The first row has to be the header separated by tab. The 'test', 'control' column are used to store expression level of individual attribute of test case and control case, respectively. The class column can be only 'gene', 'prot' and 'comp' to specify data from transcriptomic, proteomic and metabolomic data accordingly. The id column can be KEGG id with prefix or Entrez id, which correspond to the data. Conversion of Entrez id to KEGG id based on specific organism is also provided in the Ondine environment.

Statistical Global Plots and Pathway Enrichment Analysis

To provide more comprehensive view of multilevel omic data, the statistical global plot is developed. The plots are comprised of differentiate bar plots of mean expression or abundance value between test and control group, circle plots represent log fold change by circle size and log p-value by color intensity. The statistical global plots are organized based the specific pathway information. Based on a prior pi statistical value [8] cut-off, chosen by users, pathway enrichment analysis of individual omics data will be performed using hypergeometric test over pathway-gene set. The results, presented as horizontal bar chart in green color shades, are included in the statistical plot panel. This will guide users to navigate into the significant pathways. Users can further explore details of experimental data on any provided pathway maps by clicking on enrichment bar corresponding to the pathway name that notice by balloon popup then the selected map will be promptly illustrated. Transcriptomic and proteomic data are fully plotted together except for metabolomic data is plotted only enrichment bar to make the whole plot comparable.

Maps for Data Integration and Navigation

Maps in KGML format from KEGG database were transformed into JSON format for further data ovelay. Recently, KGML and KGML+ formats have been developed to manage coordinate of attributes in the metabolic maps of the KEGG database. These are the only well-defined data interchange formats that enable an easy way to integrate data with the automatic map drawing protocol, and have therefore been widely used for many applications. Using the layout from KGML files, genes and proteins are both drawn in boxes, but compounds are drawn in circles. On individual map, transcriptomic, proteomic and metabolomic data are transformed to pi value and overlaid on components by KEGG id mapping with violet color gradient. Details of data on specific components will be shown in a table inside the balloon popup when mouse cursor moving over the considered component. Components without data mapping will be colored in gray by default. Neighborhood pathways names are shaded by green color gradient corresponding to scores derived from enrichment analysis. Users can also click on the pathway name then the selected map will be promptly illustrated. In the current release of Ondine, 169 metabolic pathway maps, 272 non-metabolic pathway maps from KEGG are provided for detailed navigation and evaluation.

Features comparisons with other web application

Features of the Ondine service were compared with existed web-based pathway visualization in five major categories adapted from Kono et al., 2009: f1) pathway availability, f2) data mapping and map editing, f3) service functionality and f4) service availability. Overall comparison is shown in Table 1.

Pathway availability. Visualization of omic data on metabolic pathways is important to capture overview picture of large metabolic system. However, to overlay the data on the fully detailed metabolic map can be lost some information resolution. To overcome this problems, Ondine implemented all available maps from KEGG database, both global metabolic map and specific sub maps; such as glycolysis, The TCA cycle and amino acid metabolism. All sub maps can be opened from global metabolic map and related sub maps.

Ondine

There are both species specific and non species specific metabolic maps provided by KEGG database. Information in species specific maps are based on gene annotation of those specific species. The components in the map that not related to that species are excluded. While the non species specific map, or reference maps, provides full information of the pathway based on gene orthologous group. To be more flexible, Ondine implements only non species specific metabolic map to visualize data. The components that do not exist will be rendered in gray instead of gradient color. In order to support specific organism data visualization, Ondine provides an NCBI accession id – KEGG orthologous id conversion service.

Data mapping and map editing. The main purpose of Ondine development is to visualize multi-omic data including transcriptomic, proteomic and metabolomic data simultaneously on metabolic pathways. Experimental data can be provided by users in a simple tab-separated value format (TSV). Components on the map is comprised of rectangular boxes representing genes or proteins and circle for compounds. Ondine halves a rectangular box to simultaneously overlay transcriptome on the left and proteome on the right. Metabolomic data is overlaid on circles. To make more comprehensive view of data and assist users to identify significant pathways, a global plot was developed. The global plot represents differential plot of expression value between control and treatment for each gene or protein. Fold change and significant value from statistical test are represented in circle plot above and below the differential bar. The last component in the global plot is a enrichment bar which represents gene set enrichment analysis using hypergeometric test based on KEGG pathway assisting users to identify significant pathway and link to the metabolic map for further details.

Service functionality. Ondine interface was designed to be simple using only mouse control; such as scrolling to zoom in and out, dragging to pan, mouse over and clicking to open a dialogue box for cross reference link or further information. For providing organism specific mapping, id-conversion was implemented. Users can provide NCBI accession id instead of KEGG orthologous id (ko id) which are used in KEGG reference map. NCBI accession id will be conversed to ko id by Ondine regarding to species before data mapping. This feature is very helpful for users in order to map gene data which is usually provided with NCBI accession id without conversion.

Service availability. Ondine is purely developed with JavaScript as a free web-application to be platform independence and without any installation.

Limitation. Ondine is on going project targeting to develop a comprehensive omic visualization system and metabolic map-rendering engine. Ondine do not provide map export to image file due to the limitation of requiring image conversion library on the server. We are still finding the better solution for recording the map image right from the browser without any use of server. Nevertheless, users can use screenshot software to capture the map into image format. The interface of Ondine was designed as simple as much as possible. To accomplish the design objective some functionality has to be optimistically reduced. Map editing and time series data mapping function are also not provided by Ondine.

Conclusions

Ondine is a useful web tool for visualization of high dimensional data on biological network maps as it enables easy integrated data visualization with interactive ZUI features. Other pathway maps from Wikipathway [9], Reactome [10], PID [11] database will be populate available for navigation in the future.

Availability and requirements

Project name: Ondine : Omics navigator for data integration and evaluation
Project home page: www.ondine.se
Operating system(s): independent
Programming language: Ruby, JavaScript
Other requirements: Latest version of web browser
License: -
Any restrictions to use by non-academics: no

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NP, IN and JN designed the project. NP implemented the web application. IN conceived and conducted the project. NP and IN wrote the manuscript.

Acknowledgements

The authors thank Mike Bostock for D3js library (http://d3js.org/).This project was supported by The Knut and Alice Wallenberg Foundation Chalmers Foundation and BILS: Bioinformatics Infrastructure for Life Science

References

1. Baker M: **Gene data to hit milestone**. *Nature* 2012, **487**(7407):282-283.
2. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D *et al*: **Visualization of omics data for systems biology**. *Nature methods* 2010, **7**(3 Suppl):S56-68.
3. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M: **KEGG Atlas mapping for global analysis of metabolic pathways**. *Nucleic acids research* 2008, **36**(Web Server issue):W423-426.
4. Kono N, Arakawa K, Ogawa R, Kido N, Oshita K, Ikegami K, Tamaki S, Tomita M: **Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API**. *PloS one* 2009, **4**(11):e7710.
5. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P: **iPath2.0: interactive pathway explorer**. *Nucleic acids research* 2011, **39**(Web Server issue):W412-415.

6. Latendresse M, Karp PD: **Web-based metabolic network visualization with a zooming user interface**. *BMC bioinformatics* 2011, **12**:176.

7. Bates JT, Chivian D, Arkin AP: **GLAMM: Genome-Linked Application for Metabolic Maps**. *Nucleic acids research* 2011, **39**(Web Server issue):W400-405.

8. Xiao Y, Hsiao TH, Suresh U, Chen HI, Wu X, Wolf SE, Chen Y: **A Novel Significance Score for Gene Selection and Ranking**. *Bioinformatics* 2012.

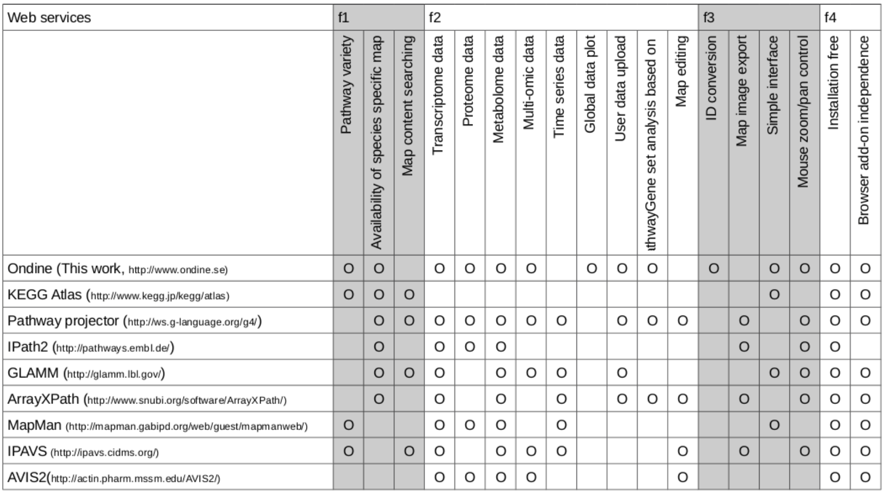
9. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR: **WikiPathways: building research communities on biological pathways**. *Nucleic acids research* 2012, **40**(Database issue):D1301-1307.

10. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR *et al*: **The Reactome pathway knowledgebase**. *Nucleic acids research* 2014, **42**(1):D472-477.

11. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database**. *Nucleic acids research* 2009, **37**(Database issue):D674-679.

Table

Table 1 Feature comparison of existing web-based metabolic map visualization and data mapping over 4 categories f1) pathway availability, f2) data mapping and map editing, f3) service functionality and f4) service availability.



Web services	f1			f2									f3				f4	
	Pathway variety	Availability of species specific map	Map content searching	Transcriptome data	Proteome data	Metabolome data	Multi-omic data	Time series data	Global data plot	User data upload	thwayGene set analysis based on	Map editing	ID conversion	Map image export	Simple interface	Mouse zoom/pan control	Installation free	Browser add-on independence
Ondine (This work, http://www.ondine.se)	O	O		O	O	O	O		O	O	O		O		O	O	O	O
KEGG Atlas (http://www.kegg.jp/kegg/atlas)	O	O	O												O		O	O
Pathway projector (http://ws.g-language.org/g4/)		O	O	O	O	O	O	O		O	O	O		O		O	O	O
IPath2 (http://pathways.embl.de/)		O		O	O	O								O		O	O	
GLAMM (http://glamm.lbl.gov/)		O	O	O		O	O	O		O					O	O	O	O
ArrayXPath (http://www.snubi.org/software/ArrayXPath/)		O		O		O		O		O	O	O		O		O	O	O
MapMan (http://mapman.gabipd.org/web/guest/mapmanweb/)	O			O	O	O		O							O		O	O
IPAVS (http://ipavs.cidms.org/)	O		O	O		O	O	O				O		O		O	O	O
AVIS2(http://actin.pharm.mssm.edu/AVIS2/)				O	O	O	O					O					O	O

Figures

Figure1. Summary of Ondine architecture is developed based on a JavaScript library built using D3js library to render SVG metabolic maps, which provide interactive responses, zoom, pan and popup window, to users.

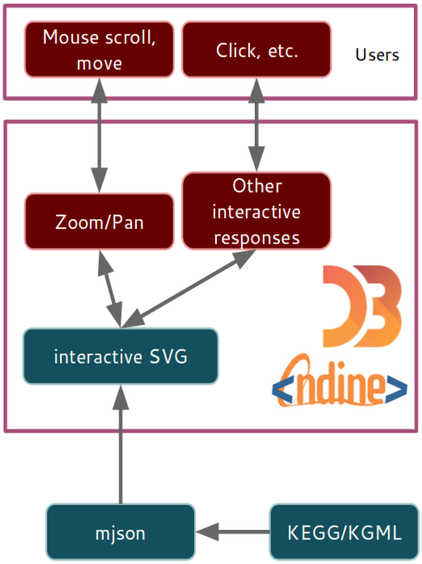
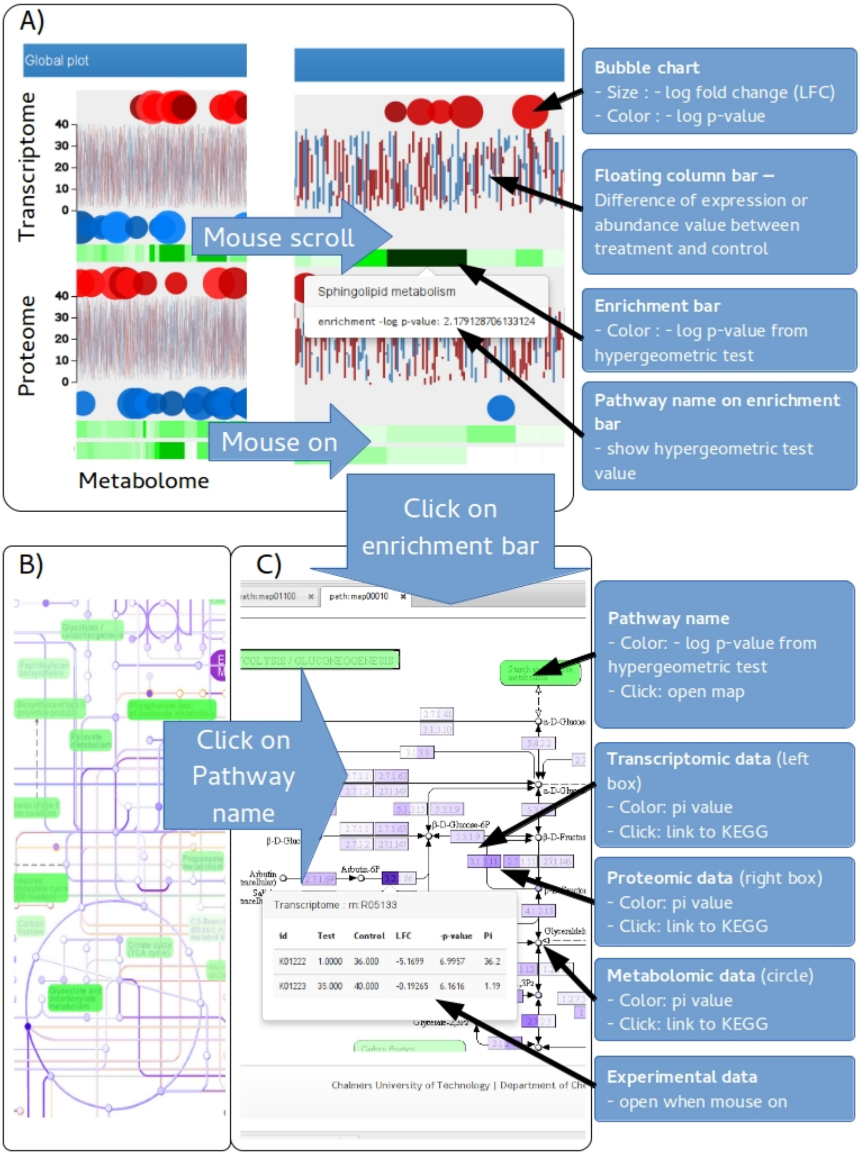


Figure2. Summary of Ondine’s interactive ZUI features for multilevel omics data integration and visualization. A) Statistical global plots, B) Global metabolic map C) Specific pathway map



Sphingolipid metabolism

enrichment -log p-value: 2.179128706133124

Transcriptome : rn:R05133

id	Test	Control	LFC	-p-value	Pi
K01222	1.0000	36.000	-5.1699	6.9957	36.2
K01223	35.000	40.000	-0.19265	6.1616	1.19

Bubble chart
- Size : - log fold change (LFC)
- Color : - log p-value

Floating column bar – Difference of expression or abundance value between treatment and control

Enrichment bar
- Color : - log p-value from hypergeometric test

Pathway name on enrichment bar
- show hypergeometric test value

Pathway name
- Color: - log p-value from hypergeometric test
- Click: open map

Transcriptomic data (left box)
- Color: pi value
- Click: link to KEGG

Proteomic data (right box)
- Color: pi value
- Click: link to KEGG

Metabolomic data (circle)
- Color: pi value
- Click: link to KEGG

Experimental data
- open when mouse on

Paper II

A dedicated database system for handling multi-level data in systems biology

A dedicated database system for handling multi-level data in systems biology

Natapol Pornputtapong*, Kwanjeera Wanichthanarak*, Avlant Nilsson, Intawat Nookeaw, Jens Nielsen[§]
Department of Chemical and Biological Engineering, Chalmers University of Technology, Göteborg, Sweden
* These authors contributed equally to this work
[§] Corresponding author
E-mail addresses:
NP: natapol@chalmers.se
KW: kwawan@chalmers.se
AN: avlant@hotmail.com
IN: intawat@chalmers.se
JN: nielsenj@chalmers.se

Abstract

Background

Advances in high-throughput technologies have enabled extensive generation of multi-level omics data. These data are crucial for systems biology research, though they are complex, heterogeneous, highly dynamic, incomplete and distributed among public databases. This leads to difficulties in data accessibility and often results in errors when data are merged and integrated from varied resources. Therefore, integration and management of systems biological data remain very challenging.

Methods

To overcome this, we designed and developed a dedicated database system that can serve and solve the vital issues in data management and hereby facilitate data integration, modeling and analysis in systems biology within a sole database. In addition, a yeast data repository was implemented as an integrated database environment which is operated by the database system. Two applications were implemented to demonstrate extensibility and utilization of the system. Both illustrate how the user can access the database via the web query function and implemented scripts. These scripts are specific for two sample cases: 1) Detecting the pheromone pathway in protein interaction networks; and 2) Finding metabolic reactions regulated by Snf1 kinase.

Results and Conclusion

In this study we present the design of database system which offers an extensible environment to efficiently capture the majority of biological entities and relations encountered in systems biology. Critical functions and control processes were designed and implemented to ensure consistent, efficient, secure and reliable transactions. The two sample cases on the yeast integrated data clearly demonstrate the value of a sole database environment for systems biology research.

Background

Systems biology aims to gain insight into complex biological systems by integrating disparate piece of data from various sources and from different levels (such as genome, transcriptome, proteome, metabolome, interactome or reactome), and formulate models that describe how the systems work [1]. The explosive growth in biological and biochemical data is beneficial for systems biology research and it has driven the development of diverse types of biological databases, such as GenBank [2], UniProt [3], SGD [4], HMDB [5], BioGRID [6], KEGG [7], ArrayExpress [8] and GEO [9]. However only 20% of the millions of deposited data in GEO have been referred in other work [9], indicating a bottleneck in utilization of large-scale data. Even though these public repositories ensure easy access to data and hence represent a platform for systems biology research, they were in many cases implemented in isolated groups with a particular purpose in mind. Furthermore, these databases often have distinct data models, different file formats, varied semantic concepts and specific data access techniques [10], and they often contain incomplete data. All in all, those factors make data management and data integration extremely challenging and error-prone.

Attempts have been made to resolve these key issues through the development of numerous data standards (e.g. SBML [11], CellML [12], PSI-MI [13], BioPAX [14], GO [15] and SBO [16]), the implementation of centralized and federated databases (e.g. cPath [17], PathCase [18] and Pathway Commons [19]) and the proposal of design methodologies for software and databases (e.g. I-cubed [20] and [21]). Although, there are still no best practices or solutions to this problem, research and development are underway by making use of current computational technologies, standards and frameworks (see [22] for a review). Here we describe the development of a dedicated database system for handling multi-level data that represents an ongoing endeavor to serve researchers in systems biology and provide alternative solutions for vital issues in data handling, data access and integration of data in a single database. The database system was designed and developed by taking into account: 1) the ability to integrate multi-level data; 2) that biological data are complex, heterogeneous, and dynamic [23]; 3) diversities of resources in terms of data model, semantic heterogeneity, data completeness and data correctness; 4) reusability, extensibility and interoperability of the system; and 5) integrity, consistency and reliability of data in the database. The design of database schema is adapted from BioPAX and implemented based on an object-oriented concept which represents practical information as an object with related attributes and a variety of relationships. This concept is applicable for biological information, which is apparently heterogeneous and sophisticated [24]. The database API was developed in C++ and included a library providing important functions to manage and interact with the system.

To illustrate the integration of multi-level data under a sole database environment, a yeast data repository was developed. The database contains multi-level data of yeast *Saccharomyces cerevisiae* (e.g. genome, annotation data, interactome and metabolic model) from different resources. Data population, data management and data access are managed by the database system. A simple query interface is provided to access the data and related information. Furthermore, two research cases were presented to demonstrate extensibility and efficiency of the database and the underlining database system in facilitating data integration tasks to achieve specific requests.

Implementation

Database system design

In order to organize complex data structure efficiently, a specific data model and management library is required to serve the bases of ACID properties including atomicity, consistency, isolation and durability to ensure the correctness of data when used. The design of the data model follows the basic concepts of a ANSI/X3/SPARC proposed architecture, which uniquely separates the view of the data structure into three layers [25]: 1) an external layer, the first layer of data abstraction in the database system, represents the entities of data to users or applications when querying; 2) a conceptual layer, the second data abstraction layer, represents entities of data that are assembled from the physical layer and are transformed to the external layer as needed; and 3) a physical layer represents the concrete data structure that is implemented in an actual file system and it is only used by the database system. These three layers are set up independently. The conceptual data structure is designed following an object-oriented data model by organizing data as a tangible object instantiated from a well-defined state, identity and behavior class [26]. In this database, class schema in conceptual layer is adapted from BioPAX ontology and implemented strictly to object-oriented concepts in a management system library to assure the accuracy and completeness of inserted data. The data schema contains 11 derived-classes as shown in Figure 1.

The database system was developed based on the MongoDB library (www.mongodb.org), thus the underlying data structure is a document-oriented data model. MongoDB was chosen on account of: 1) the database system can easily be scaled out allowing a modern data management approach such as data centric architecture can potentially be applied to this database system [27, 28]; 2) it is possible to change the data schema of the conceptual data layer implemented in the API library [27]; and 3) the MongoDB supports large file storage [28] for storing data such as gene expression data or sequencing reads. However, the schema free property of MongoDB allows storing unstructured data into the database, this might cause data inconsistency. Therefore data wrapper classes in object-oriented data model were implemented in the database API library as interface between developers to the MongoDB to ensure consistency of the stored data.

All sub-classes, implemented in the database API library, are derived from BioObject super-class with some common properties such as names, function annotation for functional ontology and cross references. The PhysicalEntity sub-classes, derived from BioObject, support molecular entities including small molecules (SmallMolecule class), DNA molecules (DNA class), genes (DNARegion class), RNA molecules (RNA class), proteins (Protein class) and molecular complex (Complex class) data. The Interaction subclasses, another BioObject derived class, support biological reactions and transport (Conversion class), molecular interactions (MolecularInteraction class), genetic interactions (GeneticInteraction class) and control interactions (Control class). Relationships among the sub-classes follow real relations of biological objects to support the data integration of multilevel data as shown in Figure 1. With this data model, reliability of data with its relationship is maintained by data classes themselves, but integrity and consistency are maintained by create, read, update and delete (CRUD) function of the library as detailed in Additional file 1.

The instances of the data classes are managed as documents classified by property “type” and pooled together in a document collection, whereas relationships between objects are separated from their own instances and pooled in another document collection to improve the efficiency of managing high complexity relationship of data. In order to optimize query time, an indexing system was applied in common query fields.

Global system architecture

The overview of the system architecture is shown in Figure 2. As the base of the system, the physical layer is managed by a document-based management system, the MongoDB, which contains the necessary interfaces; such as an interactive shell and web services. However, the MongoDB is not designed to manage a specific data structure, especially with complex relationships, and does not have features to control relationships among data objects and this may cause problems in data integrity, consistency and reliability. The database system library was therefore implemented as a core of the system, providing vital functions to manage transactions between developers and the system, and this makes it easy to populate and transform data.

Results and discussion

Applications on a yeast data repository

Given that yeast *S. cerevisiae* is a widely used model organism with abundance of genome-scale information and datasets e.g. protein-protein interactions (PPI), transcriptional regulation interactions (TRI), protein kinase interactions (KI), genome-scale metabolic model and gene annotations, integration of data from these different data sources and levels can help to gain new understanding of complex cellular systems.

We therefore developed a yeast data repository as an integrated database that contains various data of yeast. Two types of applications were built on top of the repository. One is a simple web search to query about specific biological objects. The other is additional javascripts for conducting two different research cases which utilize various data in the database to achieve the goals. Those applications are available online at http://atlas.sysbio.chalmers.se:8082. Our intention is to demonstrate efficiency of integrating data under a solitary database environment to help systems biology research rather than to present novel discoveries. As we focus on how the database system is applied, not all features of the database system are illustrated. Specific scripts were implemented to query the data stored in the database.

Data population and implementation

The yeast data repository comprises of different kinds of biological data such as genome, reactome, interactome and annotations. These data were downloaded in tab delimited or XML format from different repositories (see Table 1). The data were parsed and populated into the database using the parser library in the database system. Each biological molecule (e.g. DNA strand, gene, transcript and protein) corresponds to a specific object in the database. A unique id was assigned to each object and properties associated with it were also stored such as name, primary data source and external references.

In general, biological molecules are related to the molecule in different type (e.g. reaction performed by proteins, proteins translated from transcript, transcripts transcribed from genes and genes are on chromosome). Similar to a biological network, relationships in the database were designed in accordance with real biological phenomena. To insert an object into the database, it is required that such a relation is known. The relational reference is added together with the object and the database system will create a relation object corresponding to that relation pair. These relation objects were used in the cases below to search and explore relationship between one biological object to another.

The database system provides a practical library where each object type in the final database corresponds to a C++ object. This allows the user to fully populate the object before inserting it into the database. The database system ensures that all required data is set and pre-forms the task of inserting the object in the database. The task for the user simply becomes the task of gathering the required data, populating the object with the data and inserting the object. For each required data there exists a function such as addname and setlength to add the data to the object.

Web interface

An online web interface was developed containing links to each application: a simple query interface and a page for case demonstration. The current version allows searching for different object types such as genes, proteins, small molecules, biochemical reactions and interactions with search results that include essential objects related to the queried object. On Cases page, it comprises interactive commands used to compile the two research cases described below.

Case 1: Detecting the pheromone pathway in protein interaction networks

Signaling pathways transmit signals from one part of the cell to another part through a cascade of protein interactions and protein modifications. Cells organize cellular changes such as transcriptional programs in response to different stimuli. The yeast mitogen-activated protein kinase (MAPK) pathways are signaling pathways that have been extensively studied including pheromone response, filamentous growth, high osmolarity response and maintenance of cell wall integrity [29]. These pathways are activated by sensing stressors of protein sensors or binding of receptors to the stimuli, which in turn triggers MAPKs via a series of phosphorylations. Active MAPKs phosphorylate different targets such as protein kinases, phosphatases and transcription factors (TFs), consequently controlling cell cycle, cellular metabolism and gene expression [30]. The pheromone response pathway is activated by binding of pheromones α- and a-factor to the protein receptors Ste2 and Ste3, respectively. The signals from these membrane receptors are transmitted via sequential binding and phosphorylation reactions of MAPK cascades to TF Ste12 that subsequently activate downstream genes.

In this case, we mined the pheromone pathway segment from PPI networks where both Ste2 and Ste3 were the starting proteins and Ste12 was the ending point. However, this could result in an excessive number of candidate pathways. Several computational methods have been implemented for integrating PPIs and gene expression data or GO annotations to constraint the search [31]. To simplify the case, we integrated only PPI and GO annotation data for finding the pathways. Specific gene ontology terms (GO) in Table 2 were recursively used as constraints to eliminate proteins that are not relevant to the pheromone response pathway and consequently exclude the interactions among those proteins.

The resulting pathway contains Ste3 as a starting protein and TF Ste12 as an ending node (Figure 3). Key proteins in the pheromone pathway (e.g. Ste4, Ste5 and Fus3) were partially found in comparison to the pheromone pathway from literature [30]. This is because of: 1) the completeness of PPI data; 2) the method used for filtering unrelated proteins; and 3) the number of path lengths to search. As it is beyond our scope, we simplified the pathway construction by using only GO terms as the filters and searching with short path length (3 path lengths). The paths from Ste2 could not be identified. This is because PPI data of Ste2 could not be populated to the database. BioGRID PPI data use gene identifiers (e.g. YNR074C) for protein participants. These gene identifiers have to be mapped to UniProt protein identifiers before populating to the database. However, Ste2 gene was mapped to two proteins which conflict with data propagation rules where PPIs are represented in binary relationship. Thus whole Ste2 interactions were automatically excluded. This case highlights strict restrictions of the data population API in the database library which do not allow data population of conflict information to ensure data integrity.

Case 2: Finding metabolic reactions regulated by Snf1 kinase

Upon sensing availability of nutrients, cells undergo transcriptional, metabolic and developmental changes in order to survive under a particular nutritional state. In yeast, through complex signaling and regulatory networks, it can grow on a wide variety of nutrients e.g. glucose, galactose, glycerol and nitrogen sources. Key components in these networks include Ras/protein kinase, Snf1 and target of rapamycin complex I (TORC1) [32]. The protein kinase Snf1 is a member of the AMP-activated protein kinase (AMPK) family, which serves as a global energy regulator to ensure metabolic homeostasis of the cells. Under glucose limited condition, it allows the cells to use alternative carbon sources by regulating a set of TFs and genes in several metabolic processes including gluconeogenesis, glyoxylate cycle and β-oxidation of fatty acids [33]. In addition, Snf1 also participates in other processes such as ion homeostasis, general stress response, carnitine metabolism, pseudohyphal growth and ageing [33]. As Snf1 plays an important role in controlling many metabolic processes, we present how processes both directly and indirectly regulated by Snf1 can be retrieved from the database by integrating data from different levels.

The *SNF1* gene encodes the Snf1 catalytic subunit which regulates expression of several genes through a variety of TFs. To identify a list of metabolic reactions that are regulated by Snf1, the Snf1 protein is therefore used as a main molecule to construct the query. The first achievement is the identification of TFs, phosphorylated by the Snf1. This was done by querying for the protein targets of Snf1 from the KIs. Then with the TRIs we can retrieve the target genes of those substrates of Snf1 which are acting as TFs. From this list of the target genes we further retrieved biochemical reactions where they are involved in.

The queried result is illustrated in Figure 4. From 1333 KIs, we found Cdc14 as the substrate of Snf1 that transcriptionally regulates several metabolic genes involved in the glyoxylate cycle, amino acid biosynthesis, glycolysis / gluconeogenesis, acetate transport and oxidative phosphorylation.

Conclusions

Here we present a dedicated database model design for handling data in systems biology. It allows and supports crucial tasks in this area including integration and analysis of multi-level data, modeling of cellular pathways and collecting biological network data. In the database design, we have used a basic three layer approach to allow independent and effective implementation or changes at each data layer. The C++ library provides essential classes and services for communication among the layers. The basic properties of the database system, ACID, are responsible for providing specific functions and control processes in the library such as ”insert”, ”remove”, ”update” and ”query” to ensure that database transactions and the data inside are consistent, reliable and not corrupted. An object-oriented concept was adopted for the design and implementation of the database schema because it represents real world information as an object with related attributes and a variety of relationships. It can make the manipulation of this object and its related data easy, straightforward and relatively fast. In addition, the concept is applicable for capturing and reflecting biological information that is apparently heterogeneous and sophisticated [24]. The major design of the conceptual data structure that characterizes data in systems biology was adapted from the BioPAX ontology. Among standards, such as BioPAX, SBML and PSI-MI, for representation of biological pathway data, the main structure of them is fairly similar but BioPAX is the most general [34]. It describes biological objects in a class hierarchy, has explicit use of relations among entities and covers most of the molecular entities in biological pathways. By realizing usages of different standard formats, we included the parser classes in the library. These classes support standard formats that are generally used in most biological databases to accommodate integration of data from different sources to the database and to enhance extensibility of the system.

The database system was applied for establishing the yeast data repository, which represents an integrated platform for performing efficient systems biology research. Two applications were developed showing that building additional applications on a single database environment administrated by the dedicated database system is feasible and convenient. It should be noted that correctness and completeness of results from both research cases are not the main concern in this study, since they are depended on the quality and the availability of data sources. However the restricted control processes and functions in the database API library were designed to ensure integrity and reliability of data in the database.

We believe that the proposed database system shows an extensive attempt to serve and solve complex data handling and integration in systems biology by following and using different standards and technologies. It gives users the ability to extend and personalize the views of data through additional applications and ensures the integrity, consistency and reliability of data in the database.

Availability and requirements

- Project name: A dedicated database system for handling multi-level data in systems biology
- Project home page: http://atlas.sysbio.chalmers.se:8082
- Operating system(s): Platform independent
- Programming language: C++, php
- Other requirements: Web Browser
- Any restrictions to use by non-academics: none

Competing interests

The authors declare that they have no competing interests.

Author Contributions

NP and KW designed the database system. NP designed and coded the main library. KW designed and coded the parser class. AN populated the data and implemented the web interface. JN and IN conceived the project. NP, KW and AN wrote the paper and all authors edited it.

Acknowledgements

This project was supported by The Knut and Alice Wallenberg Foundation and the Bioinformatics Infrastructure for Life Sciences (BILS). The open access charge is supported by Chalmers Library.

References

1. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology**. *Annu Rev Genomics Hum Genet* 2001, **2**:343-372.
2. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2012, **40**(Database issue):D48-53.
3. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data**. *Database (Oxford)* 2011, **2011**:bar009.
4. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR *et al*: **Saccharomyces Genome Database: the genomics resource of budding yeast**. *Nucleic Acids Res* 2012, **40**(Database issue):D700-705.
5. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S *et al*: **HMDB: a knowledgebase for the human metabolome**. *Nucleic Acids Res* 2009, **37**(Database issue):D603-610.
6. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X *et al*: **The BioGRID Interaction Database: 2011 update**. *Nucleic Acids Res* 2011, **39**(Database issue):D698-704.
7. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 1999, **27**(1):29-34.
8. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E *et al*: **ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments**. *Nucleic Acids Res* 2011, **39**(Database issue):D1002-1004.
9. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM *et al*: **NCBI GEO: archive for functional genomics data sets--10 years on**. *Nucleic acids research* 2011, **39**(Database issue):D1005-1010.
10. Cary MP, Bader GD, Sander C: **Pathway information for systems biology**. *FEBS Lett* 2005, **579**(8):1815-1820.
11. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A *et al*: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics* 2003, **19**(4):524-531.
12. Lloyd CM, Halstead MD, Nielsen PF: **CellML: its future, present and past**. *Prog Biophys Mol Biol* 2004, **85**(2-3):433-450.
13. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C *et al*: **The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data**. *Nat Biotechnol* 2004, **22**(2):177-183.
14. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J *et al*: **The BioPAX community standard for pathway data sharing**. *Nat Biotechnol* 2010, **28**(9):935-942.
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

16. Le Novere N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM *et al*: **The Systems Biology Graphical Notation**. *Nat Biotechnol* 2009, **27**(8):735-741.

17. Cerami EG, Bader GD, Gross BE, Sander C: **cPath: open source software for collecting, storing, and querying biological pathways**. *BMC Bioinformatics* 2006, **7**:497.

18. Cakmak A, Qi X, Coskun SA, Das M, Cheng E, Cicek AE, Lai N, Ozsoyoglu G, Ozsoyoglu ZM: **PathCase-SB architecture and database design**. *BMC Syst Biol* 2011, **5**:188.

19. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C: **Pathway Commons, a web resource for biological pathway data**. *Nucleic Acids Res* 2011, **39**(Database issue):D685-690.

20. Boyle J, Cavnor C, Killcoyne S, Shmulevich I: **Systems biology driven software design for the research enterprise**. *BMC Bioinformatics* 2008, **9**:295.

21. Maier CW, Long JG, Hemminger BM, Giddings MC: **Ultra-Structure database design methodology for managing systems biology data and analyses**. *BMC Bioinformatics* 2009, **10**:254.

22. Sreenivasaiah PK, Kim do H: **Current trends and new challenges of databases and web applications for systems driven biological research**. *Front Physiol* 2010, **1**:147.

23. Ozsoyoglu ZM, Ozsoyoglu G, Nadeau J: **Genomic pathways database and biological data management**. *Anim Genet* 2006, **37 Suppl 1**:41-47.

24. Okayama T, Tamura T, Gojobori T, Tateno Y, Ikeo K, Miyazaki S, Fukami-Kobayashi K, Sugawara H: **Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library**. *Bioinformatics* 1998, **14**(6):472-478.

25. Steel BT: **Interim Report ANSI/X3/SPARC Study Group on Data Base Management Systems**. *ACM SIGMOD Record* 1975, **7**(2).

26. Hoffer JA, George, J. and Valacich, J.: **Modern Systems Analysis and Design**. In., 6 edn: Prentice Hall; 2010.

27. Quintero C, Tran K, Szewczak AA: **High-throughput quality control of DMSO acoustic dispensing using photometric dye methods**. *J Lab Autom* 2013, **18**(4):296-305.

28. Chodorow K: **MongoDB: the definitive guide**; 2013.

29. Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR *et al*: **Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles**. *Science* 2000, **287**(5454):873-880.

30. Chen RE, Thorner J: **Function and regulation in MAPK signaling pathways: lessons learned from the yeast Saccharomyces cerevisiae**. *Biochim Biophys Acta* 2007, **1773**(8):1311-1340.

31. Wang K, Hu F, Xu K, Cheng H, Jiang M, Feng R, Li J, Wen T: **CASCADE_SCAN: mining signal transduction network from high-throughput data based on steepest descent method**. *BMC Bioinformatics* 2011, **12**:164.

32. Broach JR: **Nutritional control of growth and development in yeast**. *Genetics* 2012, **192**(1):73-105.

33. Zhang J, Vaga S, Chumnanpuen P, Kumar R, Vemuri GN, Aebersold R, Nielsen J: **Mapping the interaction of Snf1 with TORC1 in Saccharomyces cerevisiae**. *Mol Syst Biol* 2011, **7**:545.
34. Stromback L, Lambrix P: **Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX**. *Bioinformatics* 2005, **21**(24):4401-4407.
35. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S *et al*: **Ensembl 2013**. *Nucleic acids research* 2013, **41**(Database issue):D48-55.
36. Osterlund T, Nookaew I, Bordel S, Nielsen J: **Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling**. *BMC systems biology* 2013, **7**:36.
37. Teixeira MC, Monteiro P, Jain P, Tenreiro S, Fernandes AR, Mira NP, Alenquer M, Freitas AT, Oliveira AL, Sa-Correia I: **The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae**. *Nucleic acids research* 2006, **34**(Database issue):D446-451.
38. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J *et al*: **Transcriptional regulatory code of a eukaryotic genome**. *Nature* 2004, **431**(7004):99-104.
39. Breitkreutz A, Choi H, Sharom JR, Boucher L, Neduva V, Larsen B, Lin ZY, Breitkreutz BJ, Stark C, Liu G *et al*: **A global protein kinase and phosphatase interaction network in yeast**. *Science* 2010, **328**(5981):1043-1046.

Figure Legends

Figure 1. Entity relationship diagram of classes and their relationship.

Figure 2. Database management system architecture.

Figure 3. Diagram of the pheromone pathway and query steps. Ovals are queried proteins in which blue ovals are included in the pathway while grey ovals are filtered out by GO terms. A green rectangular is the starting node whereas a purple diamond is the ending node where the search stops.

Figure 4. Diagram of the metabolic genes regulated by Snf1 through TF Cdc14. A green rectangular is the main gene to start the query, a blue oval is the TF and purple diamonds are metabolic genes.

Tables

Table 1 Data in Yeast multi-omic database and sources

Biological entity	Physical Entity	Data source	Amount
Chromosome	DNA	NCBI GenBank [2]	17
Gene	DNA region	Ensembl [35]	7126
RNA Transcript	RNA	Ensembl [35]	7126
Protein	Protein	UniProt [3]	6617
Compound	Small molecule	iTO977 [36]	484
Biochemical Reaction	Biochemical reaction	iTO977 [36]	717
Protein-Protein Interaction	Molecular interaction	BioGRID [6]	72453
Transcriptional regulation interaction	Control	YEASTRACT [37] and [38]	48548
Kinase interaction	Control	[39]	1333
Phosphorylase interaction	Control	[39]	254

Table 2 GO terms used for filtering proteins

GO ID	GO term
GO:0019236	response to pheromone
GO:0000750	pheromone-dependent signal transduction involved in conjugation with cellular fusion
GO:0000185	activation of MAPKKK activity
GO:0071508	activation of MAPK activity involved in conjugation with cellular fusion

Additional files

Additional file 1. CRUD functions. Details of Create, read, update and delete (CRUD) function implemented in the system library.
Format: PDF Size: 25.4KB

Additional file 2. Flow of activities in each function: A) Create; B) Delete; C) Update; and D) Read
Format: PDF Size: 0.98MB

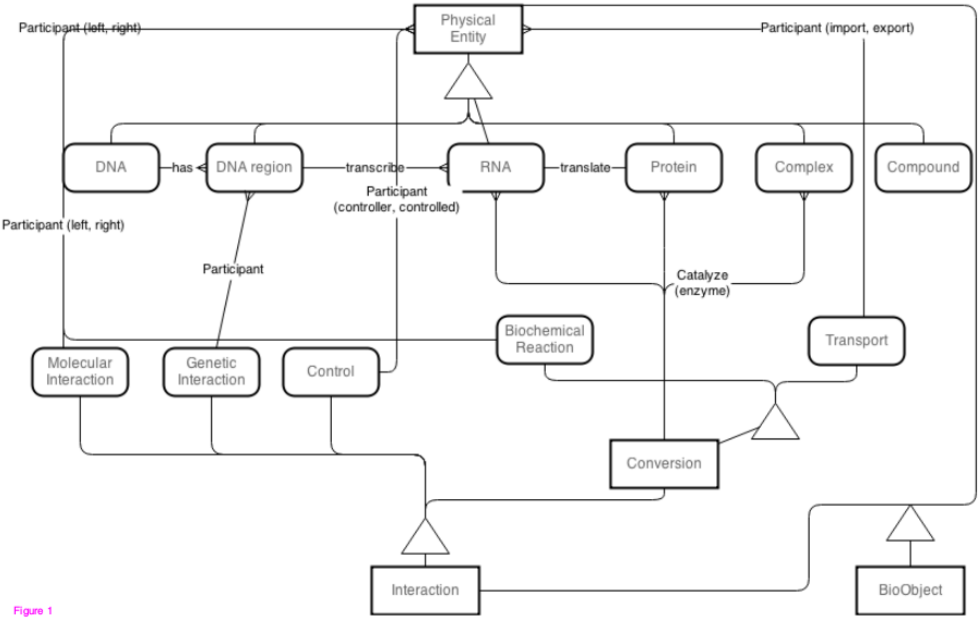

Physical Entity
Participant (left, right)
Participant (import, export)
DNA
has
DNA region
transcribe
RNA
translate
Protein
Complex
Compound
Participant (controller, controlled)
Participant (left, right)
Participant
Catalyze (enzyme)
Biochemical Reaction
Transport
Molecular Interaction
Genetic Interaction
Control
Conversion
Interaction
BioObject

Figure 1

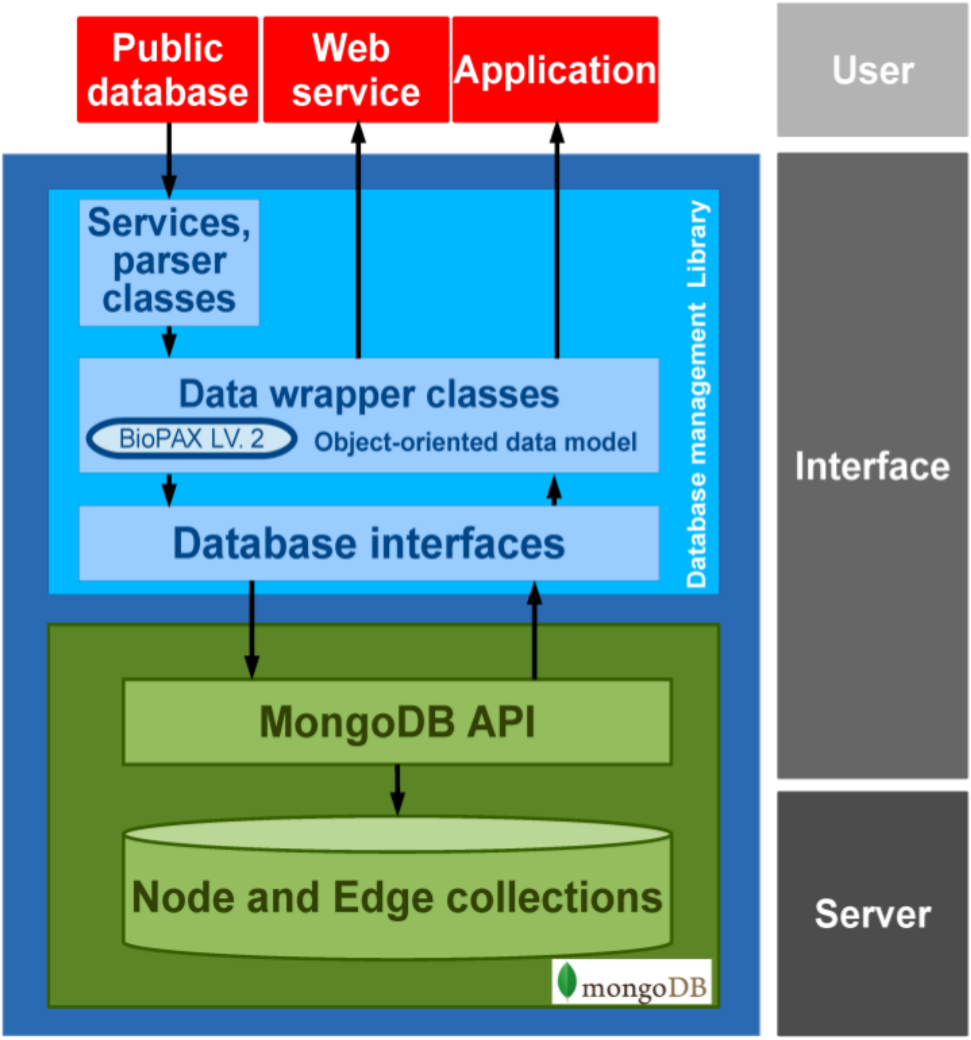
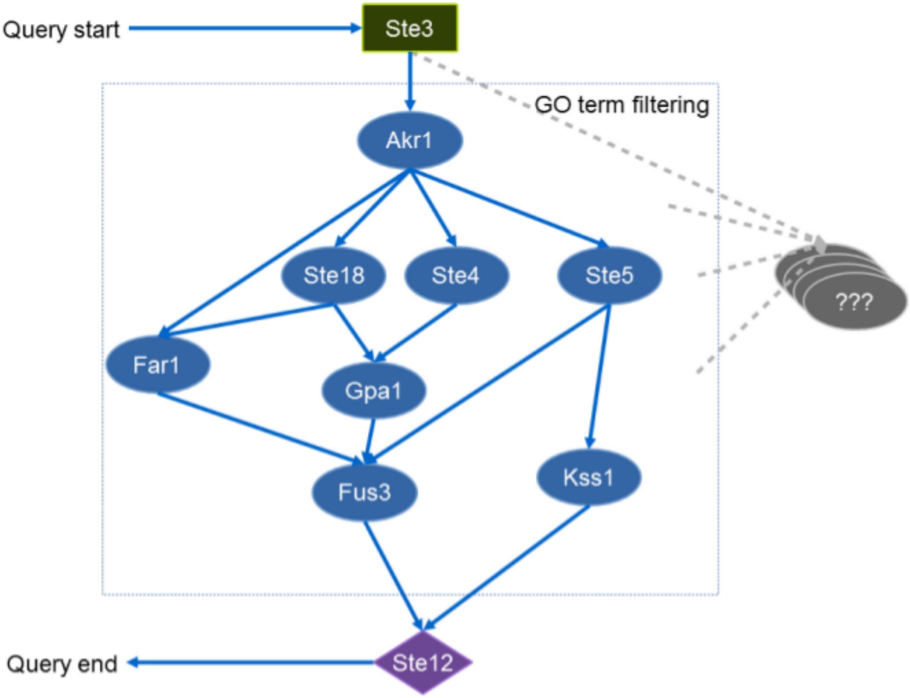

Public database
Web service
Application
User
Services, parser classes
Data wrapper classes
BioPAX LV. 2
Object-oriented data model
Database interfaces
Database management Library
Interface
MongoDB API
Node and Edge collections
mongoDB
Server

Figure 2



Query start
Ste3
GO term filtering
Akr1
Ste18
Ste4
Ste5
???
Far1
Gpa1
Fus3
Kss1
Query end
Ste12

Figure 3

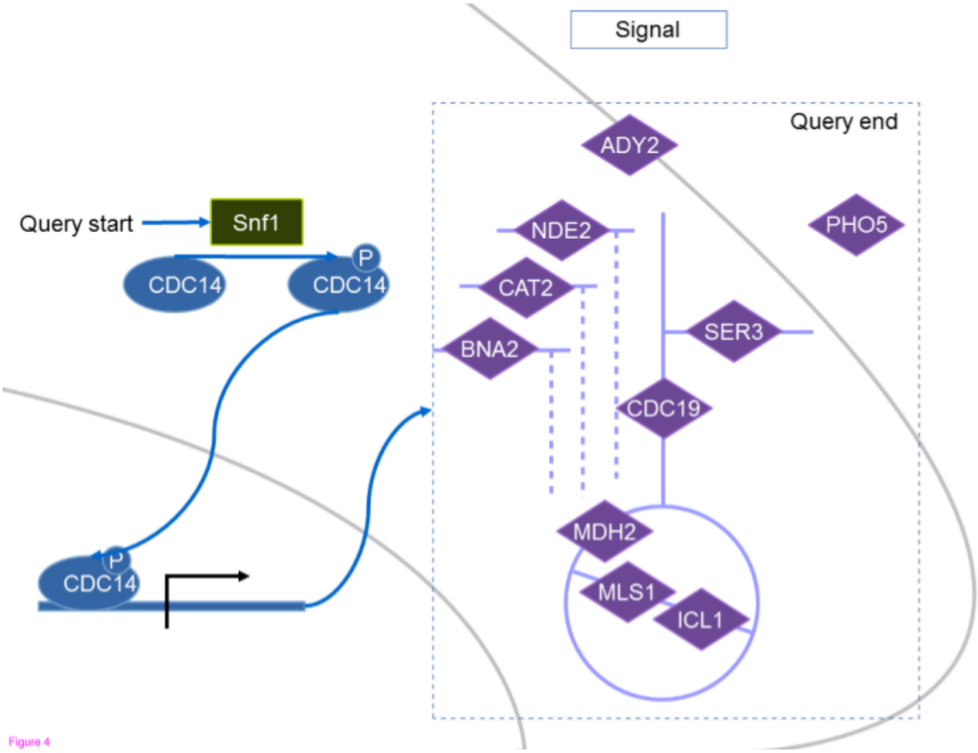

Signal
Query end
Query start
Snf1
CDC14
P
CDC14
ADY2
NDE2
PHO5
CAT2
SER3
BNA2
CDC19
P
CDC14
MDH2
MLS1
ICL1

Figure 4

Paper III

Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT

Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT

Rasmus Agren[9], Sergio Bordel[9], Adil Mardinoglu[¶], Natapol Pornputtapong[¶], Intawat Nookaew, Jens Nielsen*

Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

Abstract

Development of high throughput analytical methods has given physicians the potential access to extensive and patient-specific data sets, such as gene sequences, gene expression profiles or metabolite footprints. This opens for a new approach in health care, which is both personalized and based on system-level analysis. Genome-scale metabolic networks provide a mechanistic description of the relationships between different genes, which is valuable for the analysis and interpretation of large experimental data-sets. Here we describe the generation of genome-scale active metabolic networks for 69 different cell types and 16 cancer types using the INIT (Integrative Network Inference for Tissues) algorithm. The INIT algorithm uses cell type specific information about protein abundances contained in the Human Proteome Atlas as the main source of evidence. The generated models constitute the first step towards establishing a Human Metabolic Atlas, which will be a comprehensive description (accessible online) of the metabolism of different human cell types, and will allow for tissue-level and organism-level simulations in order to achieve a better understanding of complex diseases. A comparative analysis between the active metabolic networks of cancer types and healthy cell types allowed for identification of cancer-specific metabolic features that constitute generic potential drug targets for cancer treatment.

Citation: Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, et al. (2012) Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT. PLoS Comput Biol 8(5): e1002518. doi:10.1371/journal.pcbi.1002518

Editor: Costas D. Maranas, The Pennsylvania State University, United States of America

Received January 10, 2012; **Accepted** March 30, 2012; **Published** May 17, 2012

Copyright: © 2012 Agren et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: We acknowledge the Knut and Alice Wallenberg Foundation and the Chalmers Foundation for financial contribution to this work. The Lars Hierta Memorial Foundation supported the work on the comparative analysis of cancer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nielsenj@chalmers.se

[9] These authors contributed equally to this work.

¶ These authors also contributed equally to this work.

Introduction

Abnormal metabolic states are at the origin of many diseases such as diabetes, hypertension, hearth diseases and cancer, which can be seen in many aspects as a metabolic disease. Cancer and coronary diseases are the two main causes of death in the developed countries. It is expected that by 2030 close to 200 million persons (33% of the total population) will be obese in the EU alone, and many of these will have one or more of the following co-morbidities: diabetes, hypertension, heart disease and increased risk of cancer, and the direct (medical treatment) and indirect (inability to work) costs are estimated to amount to more than €100 billion per year [1,2]. The molecular mechanisms involved in these kinds of diseases are complex and in many cases different underlying molecular causes lead to the same disease phenotypes. A good understanding of human metabolism in different human cell types, whole tissues, and the interactions between them is therefore a necessary step towards efficient diagnosis and treatment of these diseases. Metabolism is, however, complex and involves a very large number of individual reactions that are highly interconnected through the sharing of common metabolites [3]. Understanding the function of metabolism therefore requires analysis of the complete metabolic network, and this is best done through the use of so-called genome-scale metabolic models (GEMs) [4,5,6].

There are three generic genome-scale human metabolic networks currently available, namely Recon1 [7], the Edinburgh Human Metabolic Network (EHMN) [8] and HumanCyc [9]. These reconstructions, however, are not tissue specific, which prevents their applicability to the study of particular human cell types or diseases. Tissue specific transcription profiles were used to generate tissue specific models for 10 different human tissues [10], which are subsets of Recon1, but these networks were not sufficiently flexible to explore the metabolic states of the tissues under various genetic and physiological conditions [11]. The same group later proposed a different algorithm that combines transcriptomic and proteomic data to generate a more flexible liver specific metabolic model [11], also using Recon1 as a template. Besides the mentioned automatically generated models, an extensive effort led to the publication of a manually reconstructed and annotated liver specific metabolic model referred as HepatoNet1 [12]. Models have also been developed for kidney [13], brain [14], erythrocytes [15] and alveolar macrophages [16]. Computational methods used to construct cell

Author Summary

Many serious diseases have a strong metabolic component. The abnormal metabolic states of diseased cells could therefore be targets for treatment. However, metabolism is a highly complex and interconnected system in which thousands of metabolic reactions occur simultaneously in any given cell type. In order to understand how metabolism of a diseased cell differs from its healthy counterpart we must therefore study the system as a whole. We have developed an algorithm that integrates several types of data in order to generate active metabolic networks; catalogues of the metabolic reactions that are likely to be active in a given cell type. We applied this algorithm to data for 69 healthy cell types and 16 cancer cell types. These metabolic networks can form the basis for simulation of metabolic interactions between organs or as scaffolds for interpretation of high-throughput data. We used these networks to perform an analysis between cancer and healthy cell types in order to identify cancer specific metabolic features that constitute potential drug targets. Several of the resulting targets were already known and used clinically, but we also found high-ranking reactions and metabolites which have not yet been investigated as drug targets.

type specific metabolic models aim to integrate the evidence about the presence or absence of metabolic enzymes in a particular cell type, while at the same time maintaining a well-connected network (e.g. metabolites consumed in one reaction should be able to be produced in another reaction or to be taken up from the cell environment). Transcriptome data are often noisy and differences in mRNA expression are not absolute but relative to a reference condition, and in most cases do not correlate well with enzyme levels [17]. In the frame of the Human Protein Atlas (HPA) [18,19,20] cell type specific high quality proteomic data are being generated based on specific antibodies, and this represents an essential source for protein evidence in different human cell types.

Here we present a pipeline for automatic identification of expressed cell type specific genome-scale metabolic networks (Figure 1). A key element of the pipeline is the INIT (Integrative Network Inference for Tissues) algorithm (Figure 2), which relies on the HPA as the main evidence source for assessing the presence or absence of metabolic enzymes in each of the human cell types that are present in the HPA. Tissue specific gene expression [21] was used as an extra source of evidence in INIT. Metabolomic data from the Human Metabolome Database (HMDB) [22] are also used as constraints in such a way that if a metabolite has been found in a particular tissue the resulting network should be able to produce this metabolite from simple precursors. More details can be found in the description of the method.

The output of our analysis is a cell type specific metabolic network for each of the cell types profiled in the HPA. As we are using HPA and gene expression data our networks do not represent the complete metabolic network that may be expressed in each cell type, but solely the part of the metabolic network that is expressed and hence the part of the network that is likely to be active. In order to provide a reliable and up to date genome-scale model template for our tissue/cell type specific metabolic networks, we first constructed the Human Metabolic Reaction (HMR) database containing the elements of previously published generic genome-scale human metabolic models [7,8,9] as well as the KEGG [23] database. This HMR database, which is publically available at www.metabolicatlas.com, will be periodically updated as new reactions are added to KEGG or MetaCyc or expression profiles for more proteins become available in HPA or other databases.

In order to evaluate the capability of our pipeline to generate reliable tissue specific metabolic networks, the metabolic model generated for hepatocytes was compared to HepatoNet1 [12], which is an extensively manually curated and annotated model of high quality. The availability of active metabolic networks corresponding to a broad set of healthy human cell types and cancers allows for a comparative analysis between cancer and healthy cell types in order to identify cancer specific metabolic features that constitute potential drug targets.

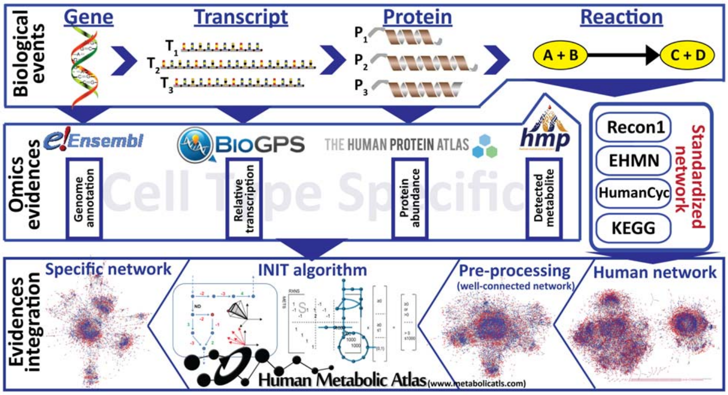


Figure 1. General pipeline used in the reconstruction of cell specific genome-scale metabolic networks. Biological information at the genome, transcriptome, proteome and metabolome levels contained in publicly available databases and generic human GEMs (Recon1, EHMN, HumanCyc) is integrated to form a generic human metabolic network, which is processed in order to obtain the connected *iHuman1512* network. Subsequently, the cell type specific evidence is used to generate cell type specific subnetworks using the INIT algorithm.
doi:10.1371/journal.pcbi.1002518.g001

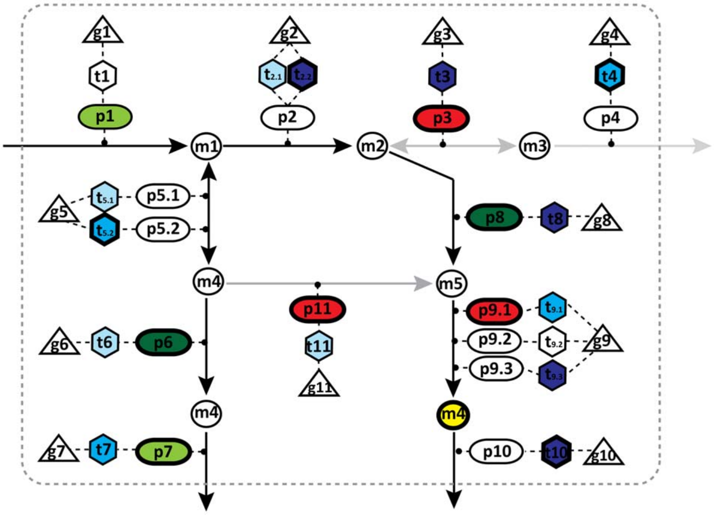


Figure 2. Illustration of the principles of the INIT algorithm. The hierarchical structure of GEMs is characterized by its gene-transcript-protein-reaction (GTPR) associations. In GEMs, each metabolic reaction is associated to one or more enzymes, which in turn are associated to transcripts and genes. Depending on the evidence for presence/absence of a given enzyme/gene in a cell type, a score can be calculated for the reaction(s) catalyzed by that enzyme. The HPA evidence scores are illustrated as red, light, medium and dark green representing negative, weak, moderate and strong evidence, respectively. The transcriptome evidence scores (GeneX), which are illustrated as red, light, medium, and dark blue representing low, medium and high expression, respectively. No evidence is present as white object. For some metabolites (yellow filled circle), metabolomic data are available to prove that they are present in the considered cell type. The aim of the algorithm is to find a sub-network in which the involved genes/proteins have strong evidence supporting their presence in the cell type under consideration. This is done by maximizing the sum of evidence scores. All the included reactions should be able to carry a flux and all the metabolites observed experimentally should be synthesized from precursors that the cell is known to take up. The bold lines represent the resulting network after optimization.
doi:10.1371/journal.pcbi.1002518.g002

Results/Discussion

Database construction

The existing methods for the inference of tissue specific active metabolic networks have only used Recon1 as a scaffold. In order to integrate other sources of information we constructed the Human Metabolic Reaction database (HMR), containing the two existing genome-scale metabolic models, Recon1 and EHMN, as well as incorporating information from HumanCyc and KEGG.

The HMR database has a hierarchical structure in which the genes are at the top and are linked to information about their tissue specific expression profiles reported by Su et al [21] via BioGPS [24]. Each gene is linked to its different splicing variants and those to their corresponding proteins. Each protein is linked to its tissue specific abundances in the HPA database [18] and to the reactions they catalyze. The reactions are linked to metabolites that themselves are linked to their tissue specific information collected from the HMDB [22]. The HMR database will be regularly updated with new reactions contained in future genome-scale human metabolic reconstructions, as well as with new evidence included in future versions of the HPA, HMDB and newly published specific transcriptome data. Details regarding the construction and curation of the HMR database are available in the Methods section. The INIT algorithm requires a connected template human metabolic model as input, and this template model was generated from HMR. The template model contains 4,137 metabolites (3,397 unique) and 5,535 reactions (4,144 unique), which are associated to 1,512 metabolic genes. This template model is referred to as *iHuman1512*.

Generation of 69 tissue specific and 16 cancer type specific genome-scale active metabolic networks

Using the INIT algorithm (see supplementary material for a detailed description), genome-scale active networks for 69 different cell types and 16 cancers were automatically generated. The resulting active metabolic networks are provided in SBML [25] format and are available at www.metabolicatlas.com.

The tissue specific models generated were compared with the BRENDA [26] collection of detected enzymes in various tissues. A hypergeometric test was carried out using the R statistical software. The reported p-values are the probabilities of obtaining an overlap higher than the observed with a random set of metabolic genes of the same size as the corresponding BRENDA entry. As it is shown in Table S1, all the comparisons between the models generated by our algorithm and BRENDA showed overlaps with p-values lower than 5e-4. Our computational liver model (*iHepatocyte1154*) shows a p-value of 1e-200, which is similar to the value obtained by comparing the manually reconstructed HepatoNet1 to BRENDA. 55% of the genes in *iHepatocyte1154* are also in BRENDA, while only 43% of the genes in HepatoNet1 are in BRENDA. The comparatively high p-values are for tissues for which there are very few annotated enzymes in BRENDA.

In order to validate the output of our algorithm, our automatically generated hepatocyte model was compared with HepatoNet1 [12], a manually curated and functional model of hepatocyte metabolism. The comparison was carried out at the gene level to avoid ambiguous decisions about reaction similarity. The overlap between the lists of genes included in each of the models is showed in Figure 3. Our hepatocyte model (*iHepatocyte1154*) contains 1,154 genes, of which 452 are also included in HepatoNet1 and 702 are absent. The evidence for the expression and translation of the 702 absent genes is as good as the evidence for the 452 genes that are in both networks, and we are therefore confident that the presence of most of the 702 extra genes has been correctly inferred by our algorithm. The HepatoNet1 network contains 261 genes not included in *iHepatocyte1154*, of which 156 were absent from our initial connected human network. Our algorithm could therefore not have assigned these genes to the hepatocyte sub-network and their existence reveals just a limitation of the data that were used as an input and not a limitation of our algorithm. 80 of these genes were not in HMR (see Table S2), but closer examination revealed that the majority (62 genes) of these genes corresponded to reactions that were actually present in HMR, but with different or absent gene associations. 13 out of the 18 remaining genes encode for transporters to the sinusoidal space; a type of blood vessels in the liver and therefore not a part of hepatocytes. The other 76 genes that were absent from *iHuman1512*, and their corresponding reactions, were removed because of being unbalanced, unconnected or otherwise problematic (see Table S3). 105 genes included in HepatoNet1, and present in *iHuman1512*, were not assigned by our algorithm to the hepatocyte-specific network. These genes correspond to 237 reactions, 132 of them still exist in *iHepatocyte1154* associated to different isoenzymes. The experimental evidences for the presence of these 105 genes (see Table S4) in the hepatocytes is mostly weak or negative, even slightly worse than the evidence for the 253 genes that were both rejected by our algorithm and absent in HepatoNet1, and we are therefore confident that these 105 genes were correctly rejected. This shows the importance of using cell type specific data when reconstructing GEMs, as enzyme isoforms can be differentially expressed in different cell types. Based on the above we can conclude that the

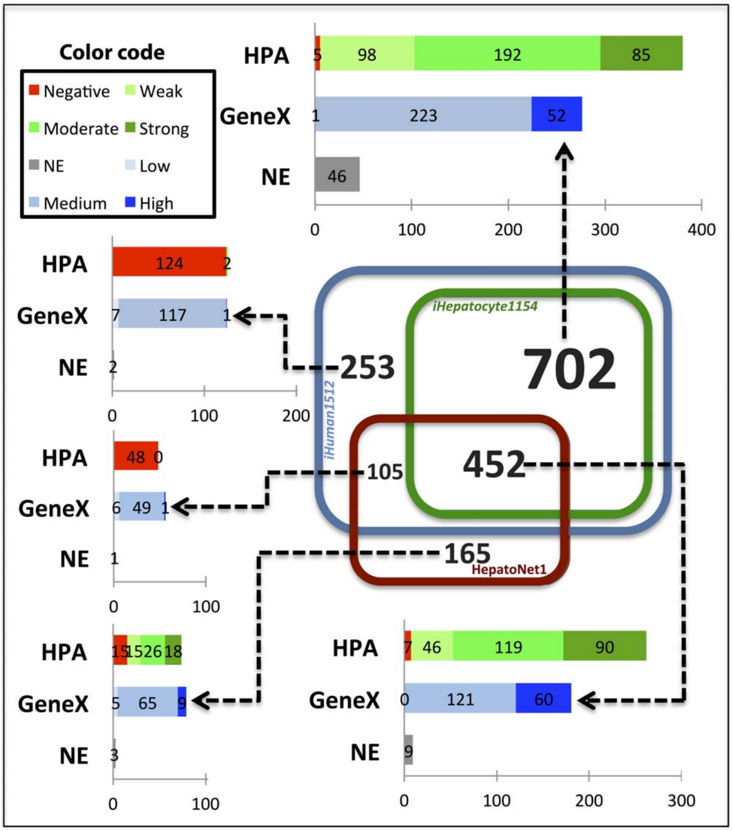


Figure 3. Gene content comparison between our hepatocyte model and HepatoNet1. The Venn diagram shows the overlap in terms of included genes between three models. The blue, green and red squares represent *iHuman1512*, our hepatocyte model *iHepatocyte1154* and HepatoNet1, respectively. The distribution of evidence scores of each section of the Venn diagram is plotted. The HPA evidence scores are illustrated as red, light, medium and dark green represent negative, weak moderate and strong expression, respectively. The transcriptome evidence scores (GeneX) are illustrated as red, light, medium and dark blue representing low, medium and high expression, respectively. No evidence (NE) is illustrated as grey color.

doi:10.1371/journal.pcbi.1002518.g003

mismatches between our hepatocyte-specific metabolic network and HepatoNet1 are accompanied by experimental evidence in favour of the choices made by our algorithm.

Finally we clustered the 69 plus 16 metabolic networks according to their similarity in terms of shared metabolic genes using unsupervised hierarchical clustering with average linkage and multiscale bootstrap resampling [27] (10,000 repetitions) implemented in the R statistical software (see Figure S1). The clustering shows, as it could be expected, local grouping of closely related cell types on the basis of cell anatomy (e.g. spleen in red pulp and white pulp cluster together). Interestingly, the cancers are separated into three different clusters, one containing liver, colorectal, breast and endometrial cancer, another minor cluster including cervical and head cancer and a third one containing the remaining ten cancers. It is also of interest to note that only 189 reactions (4.1% of the total number of reactions) are unique to a single cell or cancer type, while there is a larger core-set of 501 reactions (11.0% of total number of reactions) that are in common to all cells. Figure S2 shows the enrichment of some important metabolic pathways in the models.

Identification of cancer specific metabolic features

Since the Warburg effect was observed at the beginning of the 20th century, it is known that cancer cells show characteristic metabolic features that make them different from healthy cells [28]. This supposed metabolic similarity between cancer cells justified the development of a generic cancer genome-scale metabolic model which was used to identify potential drug targets against cancer proliferation [29]. Here we have inferred active metabolic networks for 16 different cancer types, which can be compared with the 24 healthy cell types that they come from (there are several healthy cell types for some of the tissues associated to the cancers) in order to identify metabolic features that are characteristic of cancer. A hypergeometric test was used to identify genes and reactions that tend to be present in most of the cancer specific active metabolic networks and absent in most of the original healthy cell types (see Tables S5 and S6). The p-values obtained from the hypergeometric test were used to identify Reporter Metabolites [30] that are significantly more involved in the metabolism of cancer cells (see Table S7). The sets of genes, reactions and metabolites showing enrichment in the cancer active metabolic networks with p-values lower than 1e-4 are listed in the supplementary material. These lists of genes, reactions and metabolites are cancer specific features that are likely to be playing a specific role in proliferation of cancer cells and could be potential drug targets. Our comparative analysis between two sets of active metabolic networks can be seen as a high throughput hypothesis generation method. These hypotheses are not based on mere correlations between cancer and the presence of a particular protein, but being based on the underlying metabolic network structure, and hereby our analysis provides a mechanistic interpretation about the possible role of each identified feature on the proliferation of cancer.

One of the most significant results from the Reporter Metabolites analysis is a much more pronounced metabolism of polyamines (PAs) such as spermidine, spermine, and putrescine in cancer cells. PAs play a variety of roles, of which several are related to oxidative stress prevention and suppression of necrosis [31]. PAs have long been known to be of particular importance for rapidly proliferating cells, and as such its transport and synthesis have been thoroughly investigated as anti-cancer drug targets [32]. Inhibition of single enzymes in the PA synthesis pathway has proved disappointing, due to extensive regulation of the system and use of exogenous PAs by the cancer cells. Second generation drugs instead work by targeting the transport system, by structural homology to the PAs themselves, or by linking other aneoplastic drugs to the PAs [33].

Another high-ranking target is the isoprenoid biosynthesis pathway, in particular the intermediate geranylgeranyl diphosphate. This metabolite has been shown to promote oncogenic events due to its role in prenylation of important cancer proteins such as Ras and Rho GTPases [34]. Several drugs have therefore been developed to target the prenylation process [35] or the biosynthesis of geranylgeranyl diphosphate [36].

A third prominent group among the Reporter Metabolites is prostaglandins and leukotrienes together with the intermediate HPETE. These autocrine compounds are synthesized from arachidonic acid and are elevated in connection with inflammation. They have been shown to aid in cancer progression by promoting metastasis and by influencing the immune system [37]. Of particular interest is prostaglandin E2, where both the synthesis and degradation have been investigated as promising targets for drug development [38].

The fact that so many of the identified targets correspond to well known and used drug targets, indicates that the method is able to generate biologically relevant hypotheses. Of particular interest are therefore the Reporter Metabolites that are currently not targeted in cancer treatment. Among the top-scoring Reporter Metabolites we identified biliverdin and bilirubin (Figure 4). Biliverdin reductase and the reactions catalyzed by this enzyme also appear among the genes and reaction most enriched in the cancer networks. Biliverdin reductase is known to be a major physiologic cytoprotectant against oxidative stress [39]. Cancer cells are known to be exposed to high oxidative stress resulting from the hydrogen peroxide generated during the oxidation of polyamines and other products of amino acid breakdown taking place in the peroxisome. Bilirubin is oxidized to biliverdin by hydrogen peroxide and subsequently reduced back to bilirubin by biliverdin reductase. This mechanism has been proven to be a major relief system for oxidative stress and could be considered a potential target against cancer proliferation. One of the hydrogen peroxide generating reactions taking place in the peroxisomes is the transformation of aminoacetone, which is an intermediate in the degradation of glycine, into methylglyoxal. Another source of methylglyoxal in cancer cells is from gluconeogenesis [40]. Methylglyoxal is known to be a toxic compound [41] that has been proven to induce apoptosis in some cancer cell lines [42]. Methylglyoxal also appeared among our top scoring reporter metabolites and both the gene coding for lactoylglutathione lyase (an enzyme that transforms methylglyoxal and glutathione into lactoylglutathione) and its associated reactions appear among the most enriched genes and reactions in the cancer active metabolic networks. Lactoylglutathione is further transformed into glutathione and lactic acid by the enzyme lactoylglutathione hydrolase (which also shows a significant enrichment in cancer metabolic networks with a p-value of 2e-3). Lactic acid is a well known metabolite produced by cancer cells. The mentioned two enzymes seem to be playing a relevant role in relieving the toxicity generated by methylglyoxal and could be potential drug targets against cancer proliferation. Targeting these enzymes would have the same effect on cancer cells as using methylglyoxal as a drug, but the advantage is that there would be no toxicity effects of methylglyoxal on healthy tissues.

Conclusions and perspectives

We here present a method that is able to integrate different sources of biological evidence to generate high quality cell type specific metabolic networks. We used this method to generate

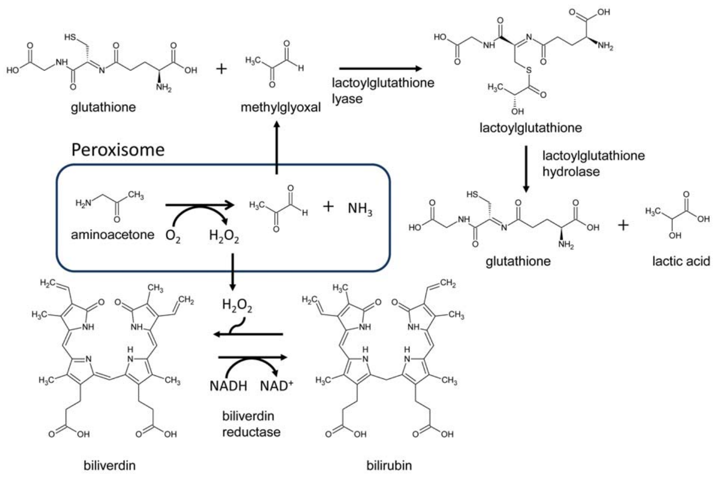


Figure 4. Example of a metabolic sub-network that was identified as being significantly more present in cancer tissues compared to their corresponding healthy tissues. Aminoacetone, which is a toxic by-product of amino acid catabolism, is converted to toxic methylglyoxal in a reaction that also result in hydrogen peroxide. The toxicity of methylglyoxal is relieved by two reaction steps involving ligation to glutathione and resulting in lactic acid. The generated hydrogen peroxide is taken care of by the enzyme biliverdin reductase. This is an example of how network-based analysis can lead to a more mechanistic interpretation of data.
doi:10.1371/journal.pcbi.1002518.g004

genome-scale metabolic networks for 69 different human cell types and 16 cancer types, and this is the first step towards the establishment of a Human Metabolic Atlas, which may become a central portal for further advancing human metabolic models with the capability of performing tissue-level and organism-level metabolic simulations, allowing for a better understanding of complex diseases. The Human Metabolic Atlas and will be made publicly accessible for the medical and scientific community and may hereby become a valuable resource in the development of personalized medicine based on system-level analysis. An example of system-level analysis is the identification of cancer specific metabolic features that we have performed by comparing the networks generated using the INIT algorithm.

Methods

Database construction

In order to have an unambiguous characterization of metabolites and reactions, KEGG and InChI identifiers were used for standardization. Metabolites lacking identifiers to external databases were left out of the HMR database together with their corresponding reactions. The metabolite identifiers were used to infer if two reactions coming from different sources were the same. Each reaction was assigned to one or several of the eight compartments included in the HMR database: nucleus, cytosol, endoplasmatic reticulum, Golgi apparatus, peroxisomes, lysosomes, mitochondria and extracellular. In cases where the subcellular localization was absent from the template models it was inferred from immunohistochemical staining in the HPA. For enzymes that were not in the HPA, Swiss-Prot and GO were used to infer localization (see Table S8 for database versions). After removing the compounds that lack identifiers, the database contained 9,922 reactions, 2,366 genes, and 9,581 metabolites for the eight different compartments (3,547 unique metabolites and 6,319 unique reactions when compartmentalization is not considered). There are 338 of these metabolites which, even if they have KEGG identifiers, are generic compounds such as "Lipid" or "2-oxoacid". Such compounds can lead to reactions that are elementally unbalanced. These problematic metabolites were removed, together with the 418 reactions in which they were involved after a detailed manual curation process. 38 reactions with wrong or unbalanced stoichiometries were also substituted by balanced versions during the curation process. In order to avoid problems associated with proton balancing, which arise from undefined protonation states of many metabolites, free exchange of protons was allowed in the models. Finally, all reactions unable to carry flux under any circumstance were removed. The reason for removing these unconnected reactions was that the algorithm requires a connected model as input. After this filtering, our template model contains 4,137 metabolites (3,397 unique) and 5,535 reactions (4,144 unique), which are associated to 1,512 metabolic genes (based on the Ensemble gene catalogue). This template model is referred to as *iHuman1512*. The numbers of genes maintained in each of the above mentioned steps are listed in Table S9. The discrepancy between the large number of reactions and the relatively small number of genes, which is also seen in previously published metabolic networks, is due to the fact that many reactions are included in the template networks based on literature studies or for connectivity reasons. In addition, some enzymes catalyze a large number of reactions and some enzymes catalyze reactions in several compartments. A comparison between *iHuman1512* and some published human metabolic networks is available in Table S10.

Algorithm for the generation of tissue-specific models

Several algorithms aiming to obtain a tissue or condition-specific active set of metabolic reactions from a generic model have been previously developed. The first of these algorithms was the Gene Inactivity Moderated by Metabolism and Expression (GIMME) algorithm [43], which uses mRNA expression data as input. Two other algorithms were developed with the specific aim of generating human tissue specific metabolic networks. The first of those [10],_ENREF_10 developed by Shlomi and co-workers, used transcriptomic data as its sole input. The second one [11] was developed by the same authors in order to obtain a functional model for human hepatocytes and is able to integrate also metabolomic and proteomic data.

The INIT (Integrative Network Inference for Tissues) algorithm is formulated as a mixed integer-linear problem (MILP) and is specially tailored to use the evidence from the HPA as input. The problem is formulated so that all reactions in the resulting model are able to carry flux. The stoichiometric matrix S contains the stoichiometric coefficients for each internal metabolite in each reaction. By multiplying the stoichiometric matrix by the vector of reaction rates we obtain a vector of net accumulation or consumption rates for each internal metabolite. Instead of imposing the steady state condition for all the internal metabolites, as it is usually done, we allow for a small positive net accumulation rate. The net productions of metabolites will be given positive weights in the optimization. The reason for this choice is that we prefer to have a network able to synthesize molecules such as NADH or NADPH, rather than only being able to use them as cofactors. If a metabolite is present in a cell type (according to the HMDB) a positive net production of this metabolite will be imposed to the network in order to assure that all the reactions necessary for its synthesis are included in the tissue specific model.

Up to this date there is not a human biomass equation available in the literature (for example Recon1 incorporates a mouse biomass equation). On the other hand, human cells (with the exception of cancer cells), in contrast to microorganisms, do not tend to proliferate or do so slowly in comparison with the rest of their metabolic functions. This makes the biomass equation less relevant, unless the aim is to model cancer proliferation. Also human cells secrete into the blood a much broader spectrum of compounds than microbial cells secrete into their environment (which are mainly fermentation products). We therefore chose to generate networks allowing for secretion (or accumulation) of all their metabolites. If we had used the stricter steady state constraint, many reactions would have been removed from the models just because they were leading to dead end metabolites. These end metabolites could in fact be added to biomass or just be secreted into the blood stream, therefore we have aimed for a more flexible approach by allowing secretion (or net accumulation) of metabolites.

The MILP used in INIT can be specified as:

$$
\begin{aligned}
&\max\left(\sum_{i\in R} w_i y_i + \sum_{j\in M} x_j\right)\\
&S\vec{v}=\vec{b}\\
&|v_i|\leq 1000y_i\\
&|v_i|+1000(1-y_i)\geq \varepsilon\\
&v_i\geq 0, i\in irreversible\ rxns\\
&b_j\leq 1000x_i\\
&b_j+1000(1-x_i)\geq \varepsilon\\
&b_j\geq 0\\
&x_j=1, j\in present\\
&y_i, x_j\in\{0,1\}
\end{aligned}
\tag{1}
$$

The parameter ε is an arbitrarily small positive number. The weights of the binary variables corresponding to the reactions account for the evidence of their presence or absence. When the corresponding enzyme has been characterized in the HPA we have used values of w_i of 20, 15, 10 and −8 for high, medium, low and absent proteins respectively. These scores are arbitrary and have been chosen to quantify the evidence colour codes that appear in the HPA. We have tested the sensitivity of the algorithm to the variation in these weights by perturbing them by 20% up and down and the impact on the output of the algorithm resulted only in small changes of the resulting networks. If the evidence comes from gene expression levels which were retrieved from BioGPS [24] and the publicly dataset "Human Body Index – Transcriptional Profiling" (GSE7307), we have used weights calculated as follows:

$$w_{i,j}=5\log\left(\frac{Signal_{i,j}}{Average_i}\right) \tag{2}$$

The signal of gene i in tissue j is divided by the average signal across all the tissues. If the signal in a particular tissue is higher than its average across all the tissues the weight will be positive, if it is lower it will have a negative weight.

For the reactions that are related to several genes or proteins the highest evidence score is used. If no gene is associated to a particular reaction, or there is no proteomic or transcriptomic evidence, a weight of −2 is used in order to avoid adding unnecessary reactions without evidence and keep the network as parsimonious as possible. If a reaction linked to several genes is added to the final tissue specific network, only the genes showing a positive evidence score are kept in the tissue specific reaction-gene association.

The MILP problem was solved using MOSEK (www.mosek.com) and its Matlab interface.

Supporting Information

Figure S1 Clustering of 69 predicted cell type specific genome-scale metabolic models for normal tissues together with 16 for cancer tissues. A dendrogram generated by unsupervised hierarchical clustering of the models based on predicted gene presence and absence is shown.
(PDF)

Figure S2 The relative pathway enrichment profiles, based on KEGG pathways, for each of the models. Blue corresponds to underrepresentation and red to overrepresentation. Note that it is the number of enzymes present for each pathway that underlie the comparison, not the abundances of the proteins.
(PDF)

Table S1 Evaluation of the models by comparison to curated tissue-specific enzymes. For each model, the set of genes is compared to the set of genes annotated as existing in the corresponding tissue in BRENDA. The p-values are derived from hypergeometric distribution.
(PDF)

Table S2 Investigation of the 80 genes that were present in HepatoNet1 but missing in HMR. The 80 missing genes were associated with 746 reactions in HepatoNet1, of which 597 metabolic and transport reactions were related to the sinusoidal space compartment. 117 metabolic reactions existed in HMR with different gene or no gene association and 32 (5 unique) reactions were altogether absent in HMR. KEGG reaction identifiers or

Transporter Classification database identifiers (TCDB) are provided for the missing reactions.
(PDF)

Table S3 Investigation of the 76 genes that were removed during the pre-processing steps. 76 genes which were present in both HepatoNet1 and the HMR database were removed in order to get a fully connected input network for the INIT algorithm. This table summarizes which genes were removed during each of the pre-processing steps (see Table S2 for details).
(PDF)

Table S4 Investigation of the 105 genes which are present in HepatoNet1 but missing in *iHepatocyte1154* due to the INIT algorithm. The 105 missing genes were associated with 182 reactions in HepatoNet1, of which 5 metabolic reactions are related to the sinusoidal space compartment. 108 metabolic reactions existed in *iHepatocyte1154* with different gene or no gene association and 69 (60 unique) reactions are absent in *iHepatocyte1154* due to INIT algorithm. KEGG reaction identifiers are provided for the missing associated reactions to the genes.
(PDF)

Table S5 List of reactions that were significantly more present in cancer tissues compared to their corresponding normal tissues (p-value<10e-4).
(PDF)

Table S6 List of genes for which their corresponding reactions were significantly more present in cancer tissues compared to their corresponding normal tissues (p-value<10e-4).
(PDF)

Table S7 List of Reporter Metabolites (p-value<10e-4).
(PDF)

Table S8 Versions of the databases used in the creation of the Human Metabolic Reaction database (HMR).
(PDF)

Table S9 Number of the genes after each pre-processing step during the generation of *iHuman1512*. Since the INIT algorithm provides a connected and functional model, reactions that are unconnected in the template model can never be included. In order to separate between reactions that were excluded due to connectivity reasons and those that were excluded due to negative evidence, a number of preprocessing steps were performed on the data in the HMR database. In the first step reactions that contain very generic metabolites such as ''lipid'' or ''alcohol'' were removed. Reactions that were elementally unbalanced were fixed or removed. Simulations were performed to ensure that the network could not gain carbon, energy or redox power in an unbalanced manner. In the second step reactions where directionality information was lacking were removed. In the third step reactions where one or more of the substrates could not be synthesized through some other reaction (unconnected reactions) were removed. Finally, in the fourth step reactions that couldn't carry flux when the model had access to all exchange metabolites (as defined in the EHMN) were removed. Consequently, *iHuman1512*, a connected human network with 5,535 reactions, associated with 1512 protein coding genes, was generated. The number of genes associated to the remaining reactions after each pre-processing steps are presented below.
(PDF)

Table S10 Comparison between iHuman1512 and some other published human metabolic networks.
(PDF)

Acknowledgments

We acknowledge valuable discussions with Prof. Bernhard Palsson and Prof. Eytan Ruppin.

Author Contributions

Wrote the paper: RA SB IN JN. Conceived the project: JN. Supervised the work: SB JN. Developed the INIT algorithm: RA. Performed the comparative analysis between cancers and healthy cell types: SB. Performed the comparison of generated models with previously published models: AM. Developed the HMR database and generated the website for the Human Metabolic Atlas: NP. Performed analysis of the generated cell type models: IN. Edited the paper: RA SB AM NP IN JN.

References

1. Caveney E, Caveney BJ, Somaratne R, Turner JR, Gourgiotis L (2011) Pharmaceutical interventions for obesity: a public health perspective. Diabetes Obes Metab 13: 490–497.
2. Rokholm B, Baker JL, Sorensen TI (2010) The levelling off of the obesity epidemic since the year 1999–a review of evidence and perspectives. Obes Rev 11: 835–846.
3. Nielsen J (2009) Systems biology of lipid metabolism: from yeast to human. FEBS Lett 583: 3905–3913.
4. Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5: 93–121.
5. Feist AM, Palsson BO (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol 26: 659–667.
6. Osterlund T, Nookaew I, Nielsen J (2011) Fifteen years of large scale metabolic modeling of yeast: Developments and impacts. Biotechnol Adv E-pub ahead of print.
7. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 104: 1777–1782.
8. Hao T, Ma HW, Zhao XM, Goryanin I (2010) Compartmentalization of the Edinburgh Human Metabolic Network. BMC Bioinformatics 11: 393.
9. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, et al. (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6: R2.
10. Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. Nat Biotechnol 26: 1003–1010.
11. Jerby L, Shlomi T, Ruppin E (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Mol Syst Biol 6: 401.
12. Gille C, Bolling C, Hoppe A, Bulik S, Hoffmann S, et al. (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. Mol Syst Biol 6: 411.
13. Chang RL, Xie L, Bourne PE, Palsson BO (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model. PLoS Comput Biol 6: e1000938.
14. Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, et al. (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. Nat Biotechnol 28: 1279–1285.
15. Bordbar A, Jamshidi N, Palsson BO (2011) iAB-RBC-283: A proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states. BMC Syst Biol 5: 110.
16. Bordbar A, Lewis NE, Schellenberger J, Palsson BO, Jamshidi N (2010) Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. Mol Syst Biol 6: 422.
17. Olivares-Hernandez R, Bordel S, Nielsen J (2011) Codon usage variability determines the correlation between proteome and transcriptome fold changes. BMC Syst Biol 5: 33.
18. Berglund L, Bjorling E, Oksvold P, Fagerberg L, Asplund A, et al. (2008) A genecentric Human Protein Atlas for expression profiles based on antibodies. Mol Cell Proteomics 7: 2019–2027.
19. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, et al. (2010) Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 28: 1248–1250.
20. Uhlen M, Bjorling E, Agaton C, Szigyarto CA, Amini B, et al. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. Mol Cell Proteomics 4: 1920–1932.

21. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062–6067.
22. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, et al. (2007) HMDB: the Human Metabolome Database. Nucleic Acids Res 35: D521–526.
23. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36: D480–D484.
24. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, et al. (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 10: R130.
25. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19: 524–531.
26. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, et al. (2010) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res 39: D507–13.
27. Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22: 1540–1542.
28. Koppenol WH, Bounds PL, Dang CV (2011) Otto Warburg's contributions to current concepts of cancer metabolism. Nat Rev Cancer 11: 325–337.
29. Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, et al. (2011) Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol 7: 501.
30. Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci U S A 102: 2685–2689.
31. Eisenberg T, Knauer H, Schauer A, Buttner S, Ruckenstuhl C, et al. (2009) Induction of autophagy by spermidine promotes longevity. Nat Cell Biol 11: 1305–1314.
32. Seiler N (2003) Thirty years of polyamine-related approaches to cancer therapy. Retrospect and prospect. Part 1. Selective enzyme inhibitors. Curr Drug Targets 4: 537–564.
33. Seiler N (2003) Thirty years of polyamine-related approaches to cancer therapy. Retrospect and prospect. Part 2. Structural analogues and derivatives. Curr Drug Targets 4: 565–585.
34. Sebti SM, Hamilton AD (2000) Farnesyltransferase and geranylgeranyltransferase I inhibitors and cancer therapy: lessons from mechanism and bench-to-bedside translational studies. Oncogene 19: 6584–6593.
35. Philips MR, Cox AD (2007) Geranylgeranyltransferase I as a target for anti-cancer drugs. J Clin Invest 117: 1223–1225.
36. Dudakovic A, Tong H, Hohl RJ (2011) Geranylgeranyl diphosphate depletion inhibits breast cancer cell migration. Invest New Drugs 29: 912–920.
37. Schneider C, Pozzi A (2011) Cyclooxygenases and lipoxygenases in cancer. Cancer Metastasis Rev 30: 277–294.
38. Eruslanov E, Kaliberov S, Daurkin I, Kaliberova L, Buchsbaum D, et al. (2009) Altered expression of 15-hydroxyprostaglandin dehydrogenase in tumor-infiltrated CD11b myeloid cells: a mechanism for immune evasion in cancer. J Immunol 182: 7548–7557.
39. Baranano DE, Rao M, Ferris CD, Snyder SH (2002) Biliverdin reductase: a major physiologic cytoprotectant. Proc Natl Acad Sci U S A 99: 16093–16098.
40. Titov VN, Dmitriev LF, Krylin VA (2010) [Methylglyoxal–test for biological dysfunctions of homeostasis and endoecology, low cytosolic glucose level, and gluconeogenesis from fatty acids]. Ter Arkh 82: 71–77.
41. Kalapos MP (1994) Methylglyoxal toxicity in mammals. Toxicol Lett 73: 3–24.
42. Kang Y, Edwards LG, Thornalley PJ (1996) Effect of methylglyoxal on human leukaemia 60 cell growth: modification of DNA G1 growth arrest and induction of apoptosis. Leuk Res 20: 397–405.
43. Becker SA, Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol 4: e1000082.

Paper IV

Human Metabolic Atlas: a web resource for human metabolism

Human metabolic atlas: An online resource for human metabolism

Natapol Pornputtapong, Intawat Nookaew, and Jens Nielsen*

Department of Chemical and Biological Engineering, Chalmers University of Technology, Göteborg, Sweden

*corresponding author

email: nielsenj@chalmers.se

Fax: +46 (31) 772 38 01

Tel: +46 (31) 772 38 54

NP's email: natapol@chalmers.se, natapol.por@gmail.com
IN's email: intawat@chalmers.se
JN's email: nielsenj@chalmers.se

Abstract

In recent years human tissue specific genome-scale metabolic (h-tGEM) modeling have provided much new information about human metabolism, especially in connection with medical research. In order to efficiently manage and utilize this kind of data, we build the Human Metabolic Atlas (HMA) website as an online resource to provide comprehensive information of human metabolism, for specific metabolic network analysis as well as to communicate with the wider research community. The HMA comprises three parts, Repository, Hreed and Atlas. The Repository is a genome-scale metabolic model collection of specific human cells and human related organisms. The models are publicly available in SBML format. Besides the models, users can easily access human reaction data from Hreed. Hreed was developed based on an object-oriented graph database system for storing standardized human metabolism information from deposited models. Connecting to the database, users can use the provided web and program interfaces to easily retrieve reaction data by using specific keywords or related genes, proteins, compounds and cross references in JSON and CSV format. With the visualization system, the Atlas, a summary of the provided h-tGEMs is overlaid on KEGG metabolic pathway maps with a zoom/pan user interface. This online resource is a useful tool for studying human metabolism at the specific cell level, organ level and for the overall human body.

Introduction

Metabolism is the process where living cells perform chemical transformations with the objective to produce building blocks for macromolecules and Gibbs free energy required for supporting cellular functions. Metabolism allows organisms to maintain their homeostasis, move, grow and reproduce. Metabolism consists of a linked series of biochemical reactions and transport processes which for most part are enzyme-catalyzed. Besides production of building blocks and energy, metabolism also involves cellular signaling that enables organisms to perceive and respond to environmental changes. Due to this central function of metabolism, disruption in parts of this may cause abnormality.

Metabolic disorders are the origin of several human diseases such as diabetes, obesity, cardiovascular disease (O'Rahilly 2009) and cancer (Seyfried and Shelton 2010). Hence, knowledge about human metabolism is fundamental to understand the mechanism of metabolic diseases, and this forms the basis for treatment or prevention. Due to the complexity of metabolism, where are large number of biochemical reactions are connected through sharing of metabolites, studying metabolism at a systemic level requires appropriate computational models, not only for human metabolism in general, but also for specific cell types or tissues.

Genome-scale metabolic models (GEM) provide comprehensive overview of the genotype to phenotype relationship in living cells and hereby provide a scaffold for interpretation of high throughput omic data in the context of metabolism (Bordbar et al. 2011; Agren et al. 2012 and Seaver et al. 2012). Recently, human tissue specific genome-scale metabolic modeling (h-tGEM) has shown to provide much new information about human metabolism when they are integrated with genomic, transcriptomic, proteomic and metabolomic data (Mardinoglu et al. 2014; Gatto, Nookaew and Nielsen 2014). There are four generic human GEMs publicly available, Recon2 (Thiele et al. 2013) which is updated version of Recon1 (Duarte et al. 2007), the Edinburgh Human Metabolic Network (EHMN) (Ma et al. 2007), HumanCyc (Romeo et al. 2005) and Human Metabolic Reaction (HMR) (Argren et al. 2012), which has been updated to HMR2.0, the most comprehensive compilation of human metabolic reactions (Mardinoglu et al. 2014), as well as, several human tissue specific cell type GEMs (i.e. from Shlomi et al. 2010; Argren et al. 2012 and Wang, Eddy and Price 2012). The h-tGEMs are widely used for data analysis, data integration and simulation to fulfill knowledge about cell type specific mechanisms prior to whole organism understanding. Knowledge from model-based data analysis can lead to understanding the mechanism of metabolic diseases which can be used to identify new therapeutic agents or novel diagnostic biomarkers in the future (Ma et al. 2007; Shlomi et al. 2010; Mardinoglu et al. 2014; Gatto et al. 2014).

A GEM consists of 3 kinds of information; 1) a list of biochemical reactions, both enzymatic and spontaneous; 2) gene-reaction or protein-reaction associations; 3) a set of objective functions that represents a range of feasible metabolic states (Lee et al. 2006; Schuetz et al. 2007; Seaver et al. 2012). Therefore, GEM reconstruction can be considered as a data integration process, which integrates genomic and metabolic reaction data. Particularly to reconstruct h-tGEM, expression data is required to discard unexpressed metabolic reactions from a generic GEM, hence the data integration process for h-tGEM is more complicated. In addition, reconstruction and accurate analysis of these complex models cannot be accomplished without appropriate computational infrastructures (Bordbar et al. 2011).

Due to the characters of biological data, integration and organization of such data is challenging. The first characteristic is the dynamics of biological knowledge. Thus, the change of biological knowledge can alter biological data interpretation which affects data modeling in database design. Therefore, to maintain the integrity of information, data models have to be adaptable (Birney and Clamp, 2004). Secondly, heterogeneity of data identifiers makes data comparison laborious and prone to error propagation. To overcome this problem, standardization of data identifiers is required (Davidson, 1995). Finally, biological components in nature behave according to the basic laws of physics. Consequently, they randomly interact and make relationship with each others (Weiss, Qu and Garfinkel, 2003). These characteristics cause complexity in biological data.

To deal with these challenges, integration and organization of biological data for GEM reconstruction, there is a need for a database system that is developed for this particular purpose. In connection with reconstruction of cell type specific models we established the HMR database using MySQL (Agren et al. 2012) and this was further expanded to HMR2.0 that covers many more reactions in an Excel database (Mardinoglu et al. 2014). Despite the extensive coverage of reactions in HMR2.0 this has drawbacks in terms of query command complexity, query time and complexity of data structure. Data models of both databases were designed for use together with the INIT algorithm (Agren et al. 2012) and expression data from Human Protein Atlas (HPA) (Uhlen et al. 2010). Therefore, these two databases were lagging the possibility to fulfill information needed in other reconstruction methods or networks integration; such as evaluation of protein families in comparative genome-scale reconstruction (Pitkanen et al. 2014) and regulatory information in probabilistic integrative modeling of genome-scale metabolic and regulatory networks (Chandrasekaran and Price 2010). Furthermore, these databases do not have any user interfaces that will enable easy access for efficiently exploration of the data. We therefore designed and implemented a new database, Hreed, that provide flexibility, consistency and crucial features for human metabolism data organization.

Reconstruction of new models is well-defined as an iterative process where the models are continuously revised. The new models are usually related to the previous models. In order to efficiently manage and utilize h-tGEM and related information, the Human Metabolic Atlas (HMA) website was built as an online resource to organize and provide comprehensive human metabolic information as models and a database to support further specific analysis or modeling as well as to communicate with the wider research community. There are three major services available on the HMA. The Repository provides management of human GEMs with a crucial feature that enables keeping track of available models. Currently the repository contains 99 human cell type specific and 3 human related microbial models available to download in SBML format. To enable exploration of the provided models, a web based metabolic map visualization named Atlas, provides a visualize system to explore human tissue specific metabolic information overlaid on KEGG metabolic pathway maps with an interactive user interface. Finally, HMA contains Hreed, a reaction database, that provides high quality of structured information of metabolic reactions and related data prior to support cloud-sourcing data curation, data expansion and programming interfaces for further utilization.

Results

The first version of the HMA website was released November 2012 as a model repository with first release of 85 h-tGEMs for normal and cancer cell types (Argren et al. 2012). In 2013 the total number of unique visitors increased 87% with a 83% increase in new visitors and a 39% increase in returning visitors. The number of visit to the model download pages also increased as shown in Figure 1. With the new implemented tools Hreed and Atlas, we will significantly improve the usability of HMA to become a key resource for studying human metabolism at the specific cell level, organ level and for the overall human body.

Human specific metabolic model repository

The first part of the HMA resource, the Repository is a collection of GEMs that have been generated automatically or been manually reconstructed for a wide range of human cell types and some human related microbial cells. The repository is classified into 4 categories; INIT normal, INIT cancer, curated and related microbial model.

INIT models were automatically generated GEMs for different human cell types by combing a generic GEM with transcriptomic and proteomic data. The first release models (r1) were generated for 69 different human normal cell types and for 16 human cancer cell types (Agren et al.2012), whereas the second release models were generated by a revised INIT algorithm for 82 different human normal cell types (Mardinoglu et al. 2014). In addition, in a study of clear cell renal carcinomas GEMs were generated for five different human cancers (r3) by the revised INIT algorithm (Gatto, Nookaew and Nielsen 2014).

Separated from INIT models, there are 2 manually curated models deposited in the GEM repository. The Adipocytes1890 model was curated by using 7,340 adipocyte associated genes from immunohistochemistry assay to be a comprehensive and functional gGEMs (Mardinoglu et al. 2013). Furthermore, to have a functional hepatocyte model for a study of non-alcoholic fatty liver disease, an extensive description of lipid metabolism from HMR2.0 was used for improving the INIT hepatocyte model resulting in iHepatocytes2322 (Mardinoglu et al. 2014).

Several recent studies have shown that the metabolism of the gut microbiome has influence on overall human metabolism, and we therefore also started to collect GEMs for human gut microorganisms. In this version of the HMA there are GEMs for three key species that are representatives of the human gut microflora (Shoai et al., 2013). These GEMs provide significant information to understand the gut microbiome ecosystem as well as its effect on human metabolism.

Hreed: Human REaction Entities Database

Hreed is a comprehensive information ecosystem that can support model reconstruction, cloud-sourcing data curation and serve as a context-aware data query system. This database is an improved version of the HMR database presented by Agren et al. 2012, and the HMR2.0 database by Mardinoglu et al. 2014. Due to the limitation of data models in the previous database, a new database system, which is designed based on an object-oriented graph data model, was implemented to manage gene, transcript, protein, small molecule and reaction data and to provide high accuracy of human metabolism information to the research community.

The data in Hreed comprises several kinds of data including human genes, transcripts, proteins, small molecules and reaction data as shown in Table 2, and these are stored in the database as instances of classes corresponding to their tangible biological entities. Genes and transcripts data were automatically populated from Ensembl 69 through the Ensembl biomart service (Flicek et al. 2013). Proteins were from Uniprot release 2012_09 (The UniProt Consortium 2013). Small molecules and reactions, which are the key data of Hreed, were parsed from HMR1.0 (Argren et al. 2012) including gene-reaction relationship information. However, to support the growth of the human metabolism information in the future, small molecule data was expanded by incorporating an internal compound data pool propagated from HMDB (Wishart et al. 2007), LMSD (Sud et al. 2007), ChEBI (Hastings et al. 2013), KEGG (Kanehisa et al. 2012) and PubChem (Bolton et al. 2014) with full InChI annotation. In addition, protein families from PANTHER (Mi et al. 2003) were populated to Hreed to support future GEM reconstructions that may involve assigning functions to unknown protein sequences.

Information in Hreed can be accessed using the graphical data query interface or the database application programming interface. The graphical data query interface, as shown in Figure 2, is suitable for general users without programming knowledge. The data objects can be queried using keywords from their properties such as names, id, cross references or other object specific properties including InChI and InChIKey. To further data searching through data relationship can be performed by specifying relation types. Query response is returned in the web as a tables by default, which provides links for user to explore further details of data from Hreed database itself or from external databases. Besides to get query response on the web, the query system also provides options for downloading data as a text file in table, XML or JSON format, which is more convenient for using data for further computational analysis.

Apart from that graphical data query system, a database API named Dactyls was implemented to support basic Create, Read, Update and Delete (CRUD) operations of the database. Besides that, a dedicated behavioral data query system based on a concept ‘Find and do’ was also implemented. The search query mimic tangible properties or function of each concrete object, which is familiar to molecular biologist. Users use a search function (beginning with ::) to retrieve object(s) first and then continuing use commands (beginning with .) to ask the user a question (ending with ?) or to do something until one reaches endpoint information as illustrated in Figure 3. After obtaining the endpoint information, users can also export gathered information to a file in JSON, XML and Table format.

Database benchmarking

For benchmarking Hreed, compound and reaction information from the global reconstruction human metabolic model, Recon version 2.02 (Thiele et al. 2013), was used for comparison. Compound data were compared based on InChI which are provided by both resources. Results in Table 3 (A) represents that 74,286 small molecules from Hreed cover more than 93% (910/974) of InChI identified compounds in Recon 2.02. Comparison of reactions from both sources using gene-reaction relationship showed that 1,162 reactions with gene association cover about 35% (881/2484) of the reactions in Recon 2.02 as shown in Table 3 (C).

Atlas: The metabolic map viewer for h-tGEMs

In order to visualize and compare h-tGEMs among different cell types, a visualization system was implemented providing a better view of tissue specific metabolic models. A web based visualization system, called Atlas, was developed with an interactive interface representing information from the latest release of INIT normal models on KEGG metabolic maps (Kanehisa et al. 2012). The Atlas is designed as a web application with a straightforward graphic user interface. The user interface comprises a pathway input box with an open button (a), tissue filter tree (b) and map viewer tabs (c) as shown in Figure 4. Users can easily open a map by typing pathway name or KEGG pathway id, which begin with 'path:map' then after by five digit of numbers in pathway input box and then click 'open' button or enter. There is also another way to open a new map by clicking on the corresponding pathway name in a map. To overlay maps with selected tissue information, the filter can be changed by unchecking the checkboxes. Only checked tissue information will be overlaid on maps. Changing the filter effects only on the map opened after changing.

The interactive interface of Atlas allows users to interactively explore the map, component details and overlaid data by using simple mouse controls. By mouse scrolling, the user can zoom in and out. Details of map components including id, name, link and graphs of overlaid data will be shown in a balloon popup after left mouse clicking on component. By clicking on reactions, usually represented as boxes, a dialog box with information of the reaction, related genes, proteins and pathways taken from KEGG, also related tissue from Hreed database, will be presented. Clicking on some texts in the dialog box will link directly to other databases; Ensembl (Flicek et al. 2013), UniProt (The UniProt Consortium 2013), KEGG (Kanehisa et al. 2012); for further information.

Discussions and conclusions

The HMA website has been developed as a comprehensive web resource to 1) provide draft h-tGEMs generated by the automatic algorithm INIT; 2) provide simulation ready functional h-tGEMs which can be used as predictive models and scaffolds for personalized genome-scale metabolic models; 3) provide tools and environment for further community driven expansion; and 4) provide visualization for comparative analysis of difference specific cell types on a metabolic map.

It has clearly been shown that to use GEMs directly as a flat file database as well as relational database to provide information of human metabolism is not efficient (Birney and Clamp 2004; Agren et al. 2012; Mardinoglu et al. 2014). The Hreed database and the Dactyls API library, developed using object-oriented graph data model, can be considered as an initial tool set to organize information about human metabolism and further application development. The database were attentively developed to support data exchange and expansion both by automatic and manual processes. In addition, The Hreed includes several data annotation schemes such as InChI, InChIKey, reaction key, MIRIAM registry and Open ontology to ensure integrity of the propagated data.

Hreed was first propagated from HMR by an automatic script. To achieve the requirements of the database system, propagated information has to be standardized. After a new cycle of model reconstruction, the curated information from GEMs will be integrated back to the database. Due to the flexible data structure, the database can be expanded further for other data categories such as complex molecule, protein-protein interaction and genetic interaction without changing the physical data layer. Besides data structure expansion ability, the database APIs also allows developers to implement the new functions; especially data integration and analysis; to the database system in the future.

Compared with the first release of the website containing only a GEM repository, the HMA has been updated with several improvements and new features, and we believe that the HMA web resource will be a good data exchange hub for the research community and facilitate new knowledge creation in human metabolism.

Methods

This section represents the details of HMA website and its tools developments.

Web site and Repository development

The HMA was developed using Ruby on Rails (RoR) as a core web application platform providing basic web application building blocks for the Repository, Hreed and Atlas development. Apart from that, RoR also supports RESTful service development which will be used for programmatic web based query services in the future.

The Repository was developed to provide GEMs of specific human tissues and related organisms. GEMs are released for downloading following publications. Recently, the models are versioned based on tissue and cell type, therefore the latest release of a specific cell type is the updated version of the same cell type model in the previous release.

Database design of Hreed

This database was designed to overcome the table flooding and low efficiency data querying that appeared in the SQL database described in Agren et al. 2012. These problems were caused by the complexity of data relationships used in modeling work, where too many data tables used in the data model causing low efficiency in the querying process and loss of accuracy in queried data. A new database management system was developed by using several data models to fit with the high complexity of data used in modeling work. The design process followed ANSI/X3/SPARC by separating the view of data structure into three layers. 1) Physical data layer represents the actual data stored in physical disk. 2) Conceptual data layer is a middle data structure representing interchange data structure between physical and external layer implemented as the core of database library. 3) external data layer represents data structure that appear to users or applications when querying (Bachman 1974). All layers were designed independently but connected and transformed by database library to let data flow through smoothly during database basic processes.

Following the database design concept, conceptual data structure was firstly designed based on combining object-oriented and graph model for separated parts of data entities and their relationship respectively. The NO-SQL data models were considered to be basis data models for new database design to avoid problem found in the SQL based data model. Object-oriented data model derived from object-oriented programming paradigm with three basis of data construction; objects, classes, attributes and relationship. Each data unit is conceptually considered as an object which can be classified into different classes by their attributes. Attributes, descriptive information of objects, are categorized into three classes of attribute: value, group (refer to a group of value or array) and aggregative (refer to another objects) attributes (Zhao 1988). However, for more flexibility to expand classes in conceptual layer and to avoid complexity of data aggregation, relationships among the classes were not designed using object-oriented data model. Relationships were modeled as objects, following a binary graph relationship model, to store connection between two objects separately from data objects. Classes, regarded as categories of objects, were modeled based on biological entity groups including DNARegion (implied gene), Transcript, Proteins, SmallMolecule and Reaction as shown in Figure 5. Classes were defined attributes limit to only closely related properties or naturally belong to class itself to avoid data redundancy and data contradiction. This minimal design limit information that can be stored in objects. To compensate for this limitation, there several functions were implemented in classes to query information, that can be implied by data context. Relationships among the classes were described as separated classed as illustrated in Figure 5 also.

The external data structure was secondly designed using the behavioral modeling concept to retrieve related information by following real object behaviors. This data structure was implemented in the query function of the Dactyls database API. The target of this querying system was to provide an easy way for biologists to get acquired data from database with less effort and without concern about the data structure. The query functions were named mimicking the actual properties or actions of real biological objects as much as possible as shown in Figure 3. However, to avoid technical problems, this work did not impact the physical data structure design. The physical data structure was defined in the underlying database management system, MongoDB.

Database and database API implementation

The design of Hreed is particularly to organize and maintain the human reaction data represented in the Human Metabolic Atlas as well as reaction related data such as genome and annotation information. In order to ensure accuracy and integrity of human reaction data, a dedicate database management system was developed using an object-oriented data model on top of a MongoDB database management system (http://www.mongodb.org/). The database management system contains a C++ library with major functions to support data manipulation including file parser, data wrapper and specific database operation functions for data population and integration.

The CRUD operations are provided in a database API library following database ACID properties to ensure accuracy and integrity of the data. Two database API libraries were developed in different programming languages and with different optimizations. The first library, Corgi, was developed in the C++ programming language, which is more efficient in speed and memory optimization, to support data modeling and database activities. Due to the nature of the C++ language, this library is quite difficult to use by non-programmers, but it is a powerful API to process the large datasets. The Corgi was used to develop the backend script of web based data query system in the Human Metabolic Atlas website. The second library is developed with Ruby scripting language called Dactyls to provide database management functions as in Corgi and a data querying system that is more readable.

Read is the most important database operation particularly for data utilization. A programmatic data querying system was developed as the external data layer by expanding data structure of conceptual data layer by using behavioral modeling concept. As it was developed in Ruby, the query system needs to run under Interactive Ruby Shell (i.e. irb or pry).

Data standardization

One of objectives of the database design is to reduce data redundancy to maintain data integrity and improve querying efficiency of the database. The uniqueness of data is also needed for data integration of multi-data level. To overcome this achievement, all data in the database must be comparable with others not only internal database but also external databases with clear identifiers. A systematic of data identifier was applied in compound, reaction data and cross-references and implemented tightly in the conceptual data layer.

Conventional chemical identifiers are normally specified by names which provides ambiguous description of chemicals. The International Chemical Identifier (InChI) and InChIKey, developed by IUPAC, is a unique computer readable identifier of chemical compounds generated from their structure (Heller et al. 2013). InChI and InChIKey is a minimum requirement for every SmallMolecule objects to put in Hreed database. MIRIAM registry, the Minimum Information Required in the Annotation of Models registry, was used and also implemented in conceptual data layers to avoid ambiguity of external id. MIRIAM registry is usually provided as URN (Unified Resources Name) composed of three parts separated by colon sign (:). The first part is prefix, always urn:miriam, to specify the source of registry from MIRIAM. The second part is a name space depended on the data collection of identifier. The last part is the identifier itself. Besides providing global unique identifiers, MIRIAM also provides a service to resolve registry back to get original data URL from original database, which is very useful to avoid dead link (Juty et al. 2012).

To make the data more consistent, identifiable, understandable and exchangeable, several data identifiers were implemented for chemical compounds, cross reference and annotation word during data propagation in the integration processes, but not for reaction data, which is a the major data used in the Human Metabolic Atlas. The specific identifier for the reaction data was developed based on the InChIKey algorithm providing 8 information layers including related compounds, reaction directionality, stoichiometry and charge balance as shown in Figure 6. Reaction key calculation was relied on InChIKey of substrates involved in the reaction. The first layer was calculated from a joined string of the first 14 characters of InChIKey of all reactants using SHA-256 hash function. To generate the string for hashing, substrates were joined together with '+' sign, as with products. Then both strings from substrates and products were joined together with '/' by alphabetical order before hashing. The second layer was also calculated as the first layer, but using the following 8 characters from InChIKey. The third layer was flagged for that whether standard InChIKeys were used in calculation or not. The fourth layer specified version of InChIKey. The fifth layer specified version of the reaction key itself. The sixth layer specified charge balance of the reaction, 'B' for balance and 'X' for not balanced. The seventh represented rounded summation of coefficient number in reaction, started with 'A' for 1 and so forth. The last layer represented the direction of reaction (Fig. 7). With this reaction key, besides to generate a unique key for reaction, users allow to use specific information layer to search for other reactions without comparing each reactants in the database.

Moreover, to support the implementation of identifier system, the database API library also provides utility functions, such as InChI and InChIKey look up services for searching the full description of compound from PubChem, a MIRIAM resolver for retrieving the original URL of specific identifier from MIRIAM registry service and a reaction key calculation function etc.

Atlas development

The Atlas was developed using the Ondine metabolic map viewer engine (Natapol Ondine paper). The engine renders the interactive SVG map using coordinate from KEGG KGML files (Kanehisa et al. 2012) which were firstly transformed into JSON format.

Availability and requirements

- Project name: Human Metabolic Atlas
- Project home page: http://www.metabolicatlas.org
- Operating system(s): Platform independent
- Programming language: C++, Ruby and JavaScript
- Other requirements: Web Browser
- Any restrictions to use by non-academics: none

Competing interests

The authors declare that they have no competing interests.

Author Contributions

NP designed the database system, web site and coded the Dactyls, Corgi library. JN and IN conceived the project. NP wrote the paper and all authors edited it.

Acknowledgements

Authors would like to thank Avlant Nilsson for web based interface data query system development and Shaghayegh Hosseini for id conversion. This project has been financially supported by The Knut and Alice Wallenberg Foundation and BILS: Bioinformatics Services to Swedish Life Science and Chalmers Foundation. The open access charge is supported by Chalmers Library.

References

Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J. 2012. Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT ed. C.D. Maranas. *PLoS Comput Biol* **8**: e1002518. http://dx.plos.org/10.1371/journal.pcbi.1002518.

Bachman CW. 1974. Summary of current work ANSI/X3/SPARC/study group: database systems. *ACM SIGMOD Rec* **6**: 16–39. http://portal.acm.org/citation.cfm?doid=983076.983078 (Accessed January 14, 2014).

Birney E, Clamp M. 2004. Biological database design and implementation. *Brief Bioinform* **5**: 31–8. http://www.ncbi.nlm.nih.gov/pubmed/15153304.

Bolton E, Wang Y, Thiessen P, Bryant S. 2008. PubChem: integrated platform of small molecules and biological activities. *Annu reports …* **4**. http://oldwww.acscomp.org/Publications/ARCC/volume4/chapter12.html (Accessed January 28, 2014).

Bordbar A, Feist AM, Usaite-Black R, Woodcock J, Palsson BO, Famili I. 2011. A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Syst Biol* **5**: 180. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3219569&tool=pmcentrez&rendertype=abstract (Accessed November 13, 2013).

Chandrasekaran S, Price ND. 2010. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* **107**: 17845–50. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2955152&tool=pmcentrez&rendertype=abstract (Accessed March 19, 2014).

Davidson S. 1995. Challenges in integrating biological data sources. *... Comput Biol* **2**: 557–572. http://online.liebertpub.com/doi/abs/10.1089/cmb.1995.2.557 (Accessed January 16, 2014).

Duarte NC, Becker S a, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ. 2007. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A* **104**: 1777–82. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1794290&tool=pmcentrez&rendertype=abstract.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41**: D48–55. http://nar.oxfordjournals.org/cgi/content/long/41/D1/D48 (Accessed January 10, 2014).

Gatto F, Nookaew I, Nielsen J. 2014. Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. *Proc Natl Acad Sci U S A* 1–10. http://www.ncbi.nlm.nih.gov/pubmed/24550497 (Accessed February 24, 2014).

Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, et al. 2013. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* **41**: D456–63. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531142&tool=pmcentrez&rendertype=abstract (Accessed January 11, 2014).

Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. 2013. InChI - the worldwide chemical structure identifier standard. *J Cheminform* **5**: 7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3599061&tool=pmcentrez&rendertype=abstract (Accessed January 13, 2014).

Juty N, Le Novère N, Laibe C. 2012. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res* **40**: D580–6. http://nar.oxfordjournals.org/cgi/content/long/40/D1/D580 (Accessed January 8, 2014).

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**: D109–14. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245020&tool=pmcentrez&rendertype=abstract (Accessed February 28, 2013).

Lee JM, Gianchandani EP, Papin J a. 2006. Flux balance analysis in the era of metabolomics. *Brief Bioinform* **7**: 140–50. http://www.ncbi.nlm.nih.gov/pubmed/16772264 (Accessed January 22, 2014).

Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I. 2007. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3**: 135. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2013923&tool=pmcentrez&rendertype=abstract (Accessed January 12, 2014).

Mardinoglu A, Agren R, Kampf C, Asplund A, Nookaew I, Jacobson P, Walley AJ, Froguel P, Carlsson LM, Uhlen M, et al. 2013. Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol Syst Biol* **9**: 649. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3619940&tool=pmcentrez&rendertype=abstract (Accessed March 19, 2014).

Mardinoglu A, Agren R, Kampf C, Asplund A, Uhlen M, Nielsen J. 2014. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat Commun* **5**: 3083. http://www.ncbi.nlm.nih.gov/pubmed/24419221 (Accessed January 22, 2014).

Mi H, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* **41**: D377–86. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531194&tool=pmcentrez&rendertype=abstract (Accessed March 19, 2014).

Pitkänen E, Jouhten P, Hou J, Syed MF, Blomberg P, Kludas J, Oja M, Holm L, Penttilä M, Rousu J, et al. 2014. Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. *PLoS Comput Biol* **10**:

e1003465.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3916221&tool=pmcentrez&rendertype=abstract (Accessed March 19, 2014).

Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD. 2005. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* **6**: R2.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=549063&tool=pmcentrez&rendertype=abstract.

Schuetz R, Kuepfer L, Sauer U. 2007. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol Syst Biol* **3**: 119.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1949037&tool=pmcentrez&rendertype=abstract (Accessed January 20, 2014).

Seaver SMD, Henry CS, Hanson AD. 2012. Frontiers in metabolic reconstruction and modeling of plant genomes. *J Exp Bot* **63**: 2247–58.
http://www.ncbi.nlm.nih.gov/pubmed/22238452 (Accessed January 27, 2014).

Seyfried TN, Shelton LM. 2010. Cancer as a metabolic disease. *Nutr Metab (Lond)* **7**: 7.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2845135&tool=pmcentrez&rendertype=abstract.

Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH, Murphy RC, Raetz CRH, Russell DW, et al. 2007. LMSD: LIPID MAPS structure database. *Nucleic Acids Res* **35**: D527–32.
http://nar.oxfordjournals.org/content/35/suppl_1/D527.abstract (Accessed January 11, 2014).

The UniProt Consortium. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* **41**: D43–7.
http://nar.oxfordjournals.org/content/41/D1/D43 (Accessed January 11, 2014).

Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, et al. 2013. A community-driven global reconstruction of human metabolism. *Nat Biotechnol* **31**: 419–25.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3856361&tool=pmcentrez&rendertype=abstract (Accessed January 9, 2014).

Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, et al. 2010. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* **28**: 1248–50.
http://www.ncbi.nlm.nih.gov/pubmed/21139605 (Accessed January 14, 2014).

Wang Y, Eddy J a, Price ND. 2012. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst Biol* **6**: 153.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3576361&tool=pmcentrez&rendertype=abstract.

Weiss JN, Qu Z, Garfinkel A. 2003. Understanding biological complexity: lessons from the past. *FASEB J* 1–6. http://www.fasebj.org/content/17/1/1.short (Accessed

March 12, 2014).

Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, et al. 2007. HMDB: the Human Metabolome Database. *Nucleic Acids Res* **35**: D521–6. http://nar.oxfordjournals.org/content/35/suppl_1/D521.abstract (Accessed January 11, 2014).

Zhao L, Roberts SA. 1988. An Object-Oriented Data Model for Database Modelling, Implementation and Access. *Comput J* **31**: 116–124. http://comjnl.oxfordjournals.org/content/31/2/116.abstract (Accessed January 12, 2014).

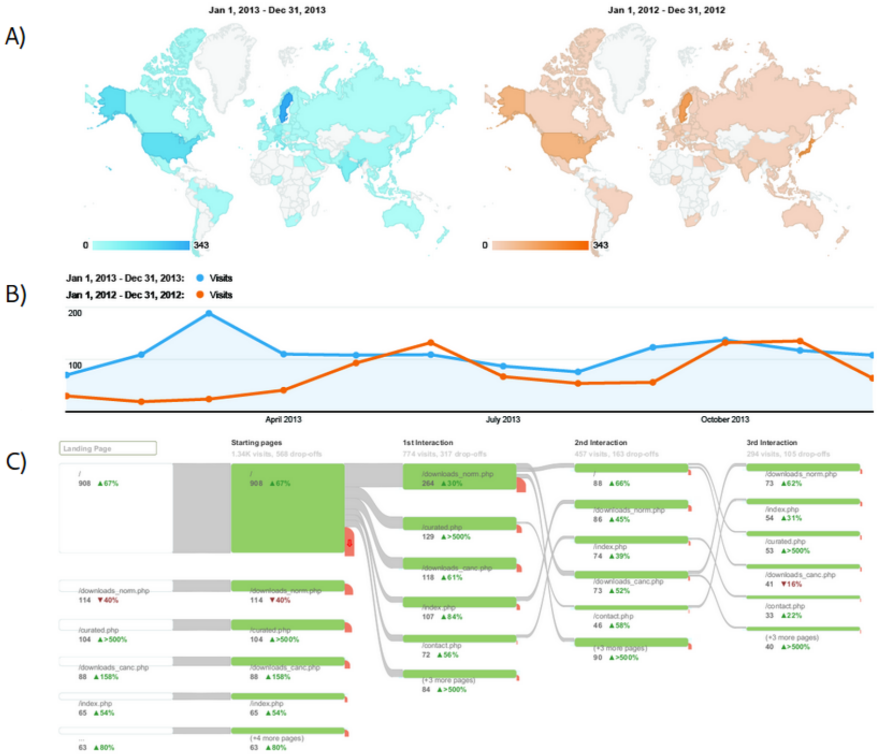
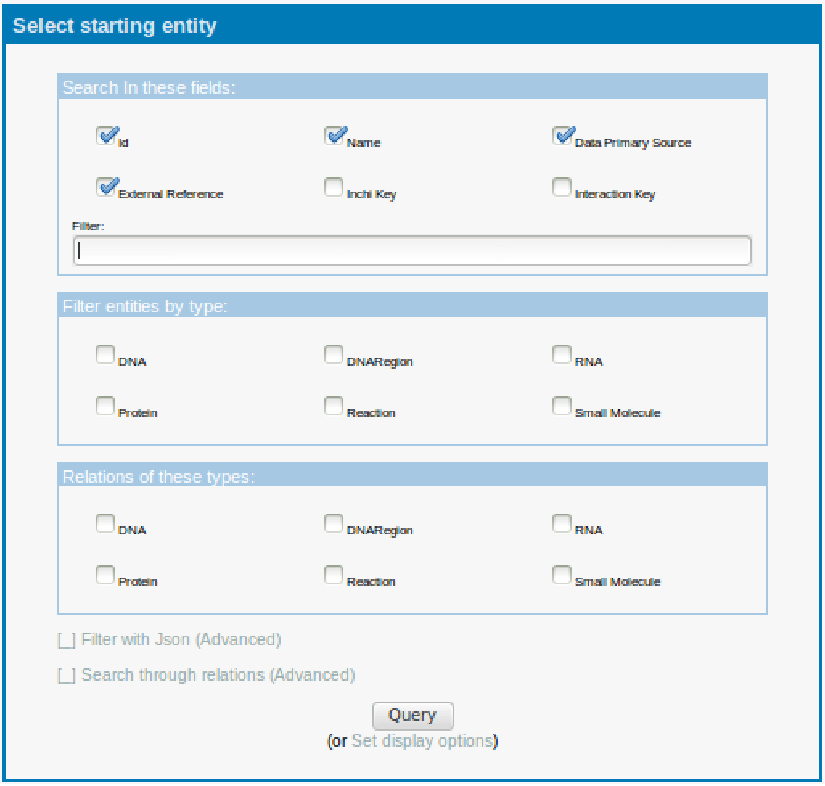


Figure 1. Visitor statistics of the Human Metabolic Atlas website generated from Google analytics comparison between year 2012 and 2013. (A) by countries (B) by months (C) by user behavior.



Select starting entity

Search In these fields:

☑ Id ☑ Name ☑ Data Primary Source

☑ External Reference □ Inchi Key □ Interaction Key

Filter:

Filter entities by type:

□ DNA □ DNARegion □ RNA

□ Protein □ Reaction □ Small Molecule

Relations of these types:

□ DNA □ DNARegion □ RNA

□ Protein □ Reaction □ Small Molecule

□ Filter with Json (Advanced)

□ Search through relations (Advanced)

Figure 2. web based data query system. Users provide keywords into the filter input box. For complex keywords, "|" and "+" can be used to combine keyword with "or" and "and" operator respectively. For searching specific object type or relationship, "filter entities by type" and "relations of these types" tab can be used.

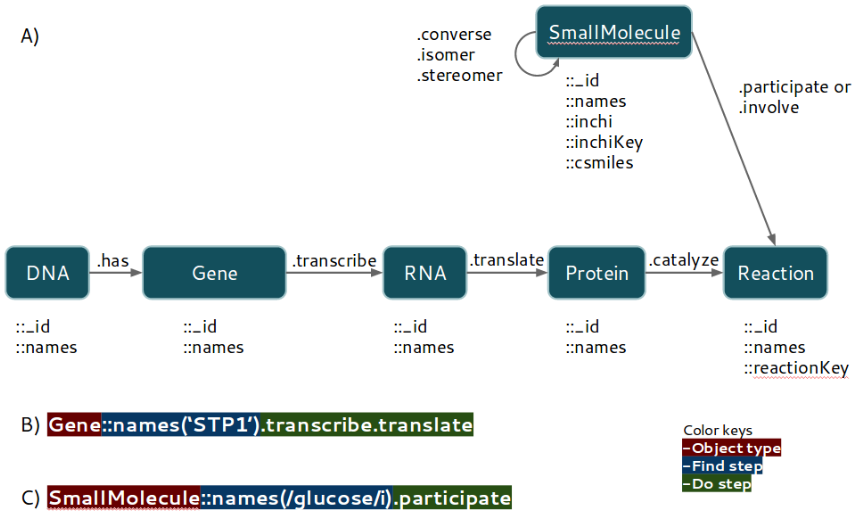
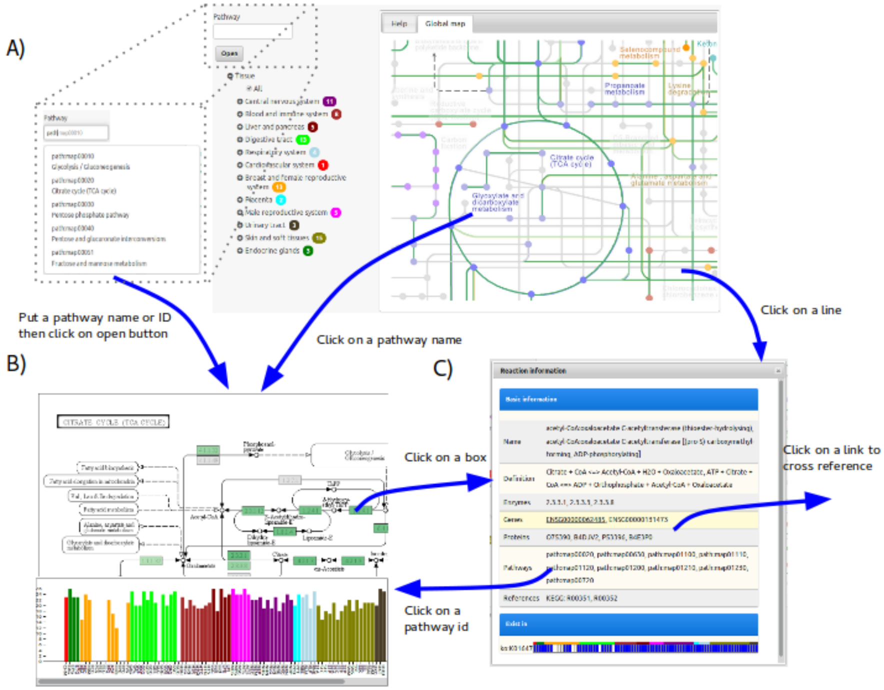


Figure 3. Programmatic data query system using Hreed (A) Object query structure with searching (begin with ::) and operating (begin with .) functions. The concepts of the system is search (for object) and operate (their functions) or search and ask (for their properties). (B) An example about searching for some proteins related to gene named 'STP1'. (C) An example about searching for reactions that catalyze compounds named with glucose.


A)
Pathway
Open
Tissue
All
Central nervous system 11
Blood and immune system 6
Liver and pancreas 5
Digestive tract 13
Respiratory system 4
Cardiovascular system 1
Breast and female reproductive system 13
Placenta 2
Male reproductive system 5
Urinary tract 3
Skin and soft tissues 15
Endocrine glands 3
pathmap00010
Glycolysis / Gluconeogenesis
pathmap00020
Citrate cycle (TCA cycle)
pathmap00030
Pentose phosphate pathway
pathmap00040
Pentose and glucuronate interconversions
pathmap00051
Fructose and mannose metabolism
Help
Global map
Selenocompound metabolism
Propanoate metabolism
Citrate cycle (TCA cycle)
Glyoxylate and dicarboxylate metabolism
Alanine, aspartate and glutamate metabolism
Put a pathway name or ID then click on open button
Click on a pathway name
Click on a line
B)
CITRATE CYCLE (TCA CYCLE)
C)
Click on a box
Click on a link to cross reference
Click on a pathway id

Reaction information

Basic information

Name	acetyl-CoA:oxaloacetate C-acetyltransferase (thioester-hydrolysing), acetyl-CoA:oxaloacetate C-acetyltransferase [(pro-S)-carboxymethyl-forming, ADP-phosphorylating]
Definition	Citrate + CoA <=> Acetyl-CoA + H2O + Oxaloacetate, ATP + Citrate + CoA <=> ADP + Orthophosphate + Acetyl-CoA + Oxaloacetate
Enzymes	2.3.3.1, 2.3.3.3, 2.3.3.8
Genes	ENSG00000062485, ENSG00000131473
Proteins	O75390, B4DJV2, P53396, B4E3P0
Pathways	pathmap00020, pathmap00630, pathmap01100, pathmap01110, pathmap01120, pathmap01200, pathmap01210, pathmap01230, pathmap00720
References	KEGG: R00351, R00352

Exist in

ko:K01647

Figure 4. Summary of Atlas map viewer functions. (A) The main view of Atlas with control panel, comprised of pathway input box for selecting specific sub metabolic map to be opened and cell type tree categorized by system for choosing cell type information to be overlaid on the map, on the left and maps on the right. Atlas starts with global metabolic pathway map by default. (B) Sub metabolic map with data overlaid and bar plot representing the number of genes that represents in this pathway map for each cell type can be opened by control panel, clicking on pathway name in every map and clicking on pathway id in information window. (C) The information window represents some information of reaction from KEGG database and provides link to external database for further information.

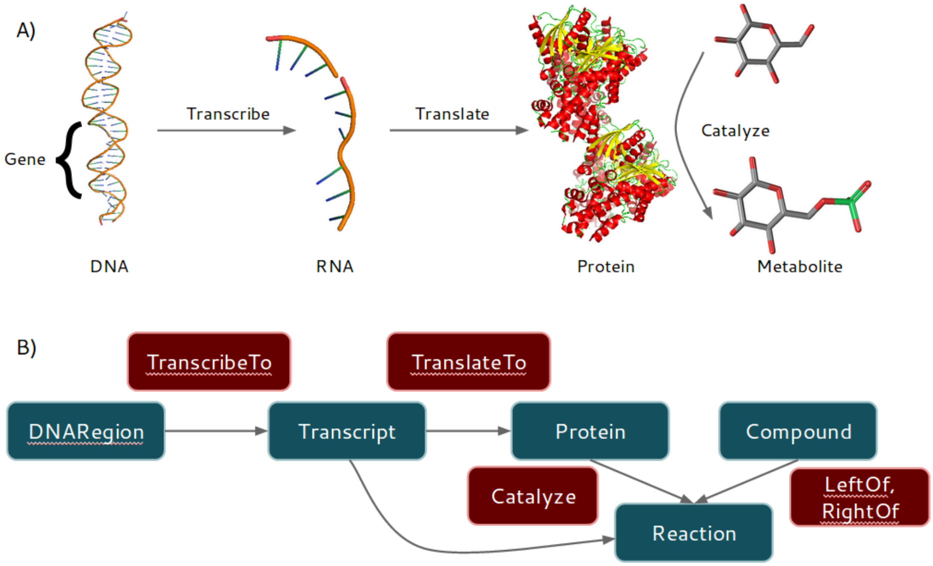


Figure 5. Data model in conceptual data layer. (The DNA, RNA and Protein molecules were generated from PDB files of 2O61, 1PNS and 3HM8 respectively using Pymol (www.pymol.org) and glucose and glucose 6 phosphate were from PubChem (Evan et al. 2014). (A) Biological entity categories and their functions. (B) Data classes reflexed to each biological entity categories and relation classes for their function.

AAAAAAAAAAAAAA-BBBBBBBBFvV-HE-D

Figure 6. Information layer of reaction key. (A) Main structures - calculated from a set of the first 14 characters of InChIKey from reactants. (B) Stereochemical information - calculated from a set of the following 9 characters of InChIKey from reactants. (F) Standard flag - S for all InChIKeys in this reaction are standard, else X. (v) InChIKey version - indicate the version of InChIKey used in calculation, if not unique X. (V) Reaction key version - A for version 1, then respectively. (H) Charge balance - B if balance, else X. (E) Stoichiometric number - Summation of all coefficient number in reaction. (D) Direction - B for backward, F for forward, R for reversible, U for unknown

Table 1. Repository summary

Human tissue specific model type	INIT normal	INIT cancer	Curated	Human related microbial
r1 (Agren et al. 2012)	69	16	-	-
r2	83	-	-	-
r3 (Gatto et al. 2014)	-	5		-
r4 (Mardinoglu et al. 2013)	-	-	1	-
r5 (Mardinoglu et al. 2014)	-	-	1	-
r6	-	-	-	3

Table 2. Summary of the Hreed database.

Datasets	Source	Data types	Imported method	Imported/Total
Ensembl gene 69	biomart.org	Gene (and Chromosome)	Automatically	62311
Ensembl transcript 69	biomart.org	Transcript	Automatically	213272
UniProt 2012_09	uniprot.org	Protein	Automatically	19084
HMR compound	metabolicatlas.org	Compound	Partial curated	1692/3539
Pooled dataset	metabolicatlas.org	Compound	Automatically	72594
HMR reaction	metabolicatlas.org	Biological reaction	Partial curated	5282/5526

Table 3. Comparison between Hreed and Recon 2.02

a) Compound

	Recon: Compound	Hreed: SmallMolecule
All	5063	74286
Unique/non-compartment	2626	74286
with InChI	1332	74286
with standard InChI	974	74286

b) Reaction

	Recon: Reaction	Hreed: Reaction
All	7440	2250
Unique/non-compartment	4910	2250
Metabolic Reaction	3309	2250
Transport Reaction	1601	0

c) Reaction based on gene association

	Recon: Reaction	Hreed: Reaction
All	4252	1162
Unique/non-compartment	2875	1162
Unique metabolic Reaction	2484	1162
Transport Reaction	391	0

Publication V

A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae