GeneReg: Integration of Experimental Data on the DNA Transcription Process [1]

David Pastor [a] Karen Lemmens [b] Álvaro Cortés-Calabuig [a]
Kathleen Marchal [c] Marc Denecker [a] Bart De Moor [b]

[a] *Department of Computer Science, Katholieke Universiteit Leuven, Belgium.*
[b] *Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium.*
[c] *Department of Microbial and Molecular Systems, Katholieke Universiteit Leuven, Belgium*

Abstract

We present *GeneReg*, an application that integrates heterogeneous databases describing the DNA transcription process of *Saccharomyces cerevisiae*. The architecture of GeneReg is based on a standard mediator-based paradigm: a global relational schema is designed as a target for information available in a number of data sources spread over the Internet. The specification of the logical relations between global and local ontologies is defined by means of simple schema matching rules. The data is then retrieved automatically from the sources –using the rules– and stored in the global schema, that in GeneReg represents the abstraction of the gene regulation process. The final product is the representation of genomic data dispersed on the web in a uniform interface. The application runs in Windows platform, it is based on MySQL 5.1 database management system and at the current stage extracts information from the Stanford microarray, ChIP-chip and Motif databases.

1 Introduction

In the last decade, the number of publicly available biological data has increased considerably. The reconstruction of genetic regulatory networks, based on those high-throughput data is one of the foremost challenges of current bioinformatics research [7, 2]. Because regulatory networks are modular and hierarchically organized [20], they can be described in terms of modules which drastically reduces the complexity of the network inference problem. A module is defined as a regulatory program and a corresponding set of co-expressed genes. The program consists of a set of regulators and their corresponding motifs.

Traditionally, module identification methods deal with each of the different data sources separately (for example, solely based on microarrays [18]). However, simultaneous analysis of distinct data sources has a major advantage over their separate analysis: their integration allows gaining holistic insight into the network and a more refined definition of transcriptional modules can be derived [23]. Therefore, more recent approaches for module inference combine several data sources [9, 12]. These algorithms all need large collections of high-throughput data (like microarray data, ChIP-chip data or motif data) as input to be able to unravel the regulatory network.

Among the facing challenges for potential users of this rich source of information is how to integrate the different available data sources. The information is not only stored in different formats and data models, but also contains redundant, incomplete and/or inconsistent data. The sources even present uneven levels of structure: relational databases, semistructured XML documents and plain HTML documents count among the available formats.

In order to address the general problem of integrating heterogeneous databases, an architecture called mediator-based system was proposed in the mid-90's (see [13] for an overview). The central idea of a mediated system is to provide users with a single interface to access the information available in the sources. The

[1] This work is based on the master thesis of the first author. Research supported by Research Foundation-Flanders (FWO-Vlaanderen) and by GOA 2003/08 "Inductive Knowledge Bases".

interface –usually referred as the global schema– consists of a number of relations describing an abstraction of the domain under consideration. The global schema then *virtually* imports the relevant information from the data sources by means of logical formulas encoding the semantic relations between the global and local schemata. Two main approaches have been proposed to specify the logical relations: *Local-as-View (LAV)* and *Global-as-View (GAV)*. In the LAV approach a data source predicate is defined as a *view* of the global predicates; the GAV is the converse, global predicates are expressed as views in terms of the local predicates (for a detailed comparison of the two approaches see [22]).

In this paper we describe GeneReg, an application that follows the mediator-based philosophy under the GAV paradigm to integrate a number of data sources conveying information about a DNA transcription process of the *Saccharomyces cerevisiae* (or yeast). Instead of retrieving the information from the data sources in each query –as in standard mediator based systems–, GeneReg actually *stores* the relevant data as an instance of the global schema. The result is then a conceptually well-founded and data-rich repository of selected information regarding the transcriptional regulatory network of the yeast. The repository can then be directly queried, data-mined or exported by a biologist.

This paper is organized as follows: in section 2 we provide the basic building block of biology and databases theory that are used in the application. In sections 3 and 4 we describe the components of GeneReg and how the integration process is carried out. In section 5 we review related work. We conclude the paper in section 6 proposing some future lines of research.

2 Preliminaries

GeneReg comprises theoretical elements from two different fields, therefore we split this section in biology and database subsections.

2.1 The Biological Domain

DNA is the cellular library that contains the necessary information to create, develop and maintain a living cell. DNA forms a double helix consisting of two chains with complementary nucleotide sequences. The DNA sequence is formed by the nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). The complementary of DNA stems from the binding of A to T and C to G [15].

While DNA contains the code of life, the proteins are the real functional elements within a cell. It is the code of these building blocks of a cell that is contained in the genes hidden in the DNA sequence. Genes are segments of DNA that encode for proteins through the intermediate action of mRNA. To form a protein, a gene is first transcribed into a RNA sequence. RNA or ribonucleic acid is a single stranded molecule similar to DNA. The transcription of a gene means that a RNA sequence complementary to one of the DNA strands is formed. This RNA sequence is further processed to form messenger RNA (mRNA), the template of the protein sequence. mRNA is translated into an amino acid sequence (or protein) using a specific genetic code. Each three nucleotides of the mRNA sequence are translated into one particular amino acid. Such a triplet of nucleotides is called a codon [15].

Besides the transcribed sequences or genes, there are also untranscribed sequences present in the DNA. These regions play a major role in the regulation of transcription. Indeed, regulation mechanisms are needed because genes will only be transcribed when the corresponding protein is needed in the cell [15].

2.1.1 Gene regulation

One of the basic principles of molecular cell biology is that the characteristics and behaviour of a cell are determined by the set of proteins contained in the cell. This set of proteins is determined by the stability of these proteins, the set of mRNAs that are translated and the translation rate of these mRNAs. The level of mRNA is, in turn, determined by the genes that are transcribed and the rate of their transcription.

Unravelling the mechanisms that regulate gene activity in an organism is a major goal of molecular biology. The focus of this manuscript will be on the regulation of transcription. The transcription process is initiated by the binding of several transcription factor proteins (also called regulators) to regulatory sites in the DNA, located in the promoter region of the gene. These binding sites or regulatory motifs are small, conserved DNA regions, and they form the tags in the DNA that are recognised by a regulator. The transcription rate can be positively or negatively affected by the action of transcription factors. When the transcription factor significantly decreases the transcription of a gene, it is called a repressor. If, on the other hand, the expression of a gene is upregulated, biologists speak of an activator.

The transcriptional regulatory network is defined by which transcription factors bind to which promoters and what the integrated effect of all these transcription factors is on the expression of all genes [8].

Large quantities of high-throughput data permit deriving full regulatory network structure based on experimental data. Different ways of combining expression data, ChIP-chip data and motif analysis data have allowed the generation of hypothetical regulatory network structures using a variety of methods in a number of model organisms like yeast [12].

2.2 Databases

We now recall some basic concepts of first-order logic and databases theory. We consider a first-order vocabulary Σ consisting of predicate symbols $\mathcal{R}(\Sigma)$ and an infinite set of constants $\mathcal{C}(\Sigma)$. Atomic formulas are constructed from the predicates in $\mathcal{R}(\Sigma)$ over tuples $\bar{t}$ of constants from $\mathcal{C}(\Sigma)$. First-order formulas over Σ are constructed from the atomic formulas using the standard rules for $\neg, \wedge, \vee, \forall, \exists$. An integrity constraint $\mathcal{I}$ is a first-order formula over Σ without free variables.

A first-order structure for Σ (sometimes called Σ-interpretation) is a triple $I = \langle Dom^I, C^I, R^I \rangle$, where Dom^I is a non-empty set called the domain of discourse of I, and C^I and R^I are functions interpreting the constants and the predicate symbols of Σ.

Definition 1 *A data source is a pair $S = \langle \Sigma_S, D \rangle$, where Σ_S is the vocabulary of the data source and D is a finite set of atoms constructed with predicate symbols from $\mathcal{R}(\Sigma_S)$ and the constants in $\mathcal{C}(\Sigma_S)$.*

We adopt an open-world assumption for data sources, i.e. each data source is a partial –but sound– description of the domain of discourse. It is also assumed that the unique name axioms hold. The semantics of a data source is then defined as a first-order theory based on Σ_S.

Definition 2 *The semantics of a data source $S = \langle \Sigma_S, D \rangle$ is given by the theory consisting of the following axioms:*

Soundness*:* $\bigwedge_{A \in D} A$

Unique Name Axiom *(*$\mathsf{UNA}(D)$*):* $\bigwedge_{1 \leqslant i < j \leqslant n} C_i \neq C_j$

where $C_1, \ldots, C_n$ are the constant occurring in D.

Example 1 *The following data source $S = \langle \Sigma_S, D \rangle$ stores information about gene names and their IDs.*

$$D = \{GeneName(1a, TFC3), GeneName(2a, EFB1), GeneName(3b, KRH1)\}.$$

The vocabulary Σ_S consists of the predicate symbol $GeneName/2$ and includes all constants appearing in D.

For the next definition, we consider only one predicate symbol per data source. This assumption does not harm generality since it is always possible to view a data source with multiple predicate symbols as the union of sources consisting of a single predicate.

We are now in conditions to formally define a data-integration system [2].

Definition 3 *A* data-integration system *based on a relational vocabulary $\Sigma = \Sigma_G \cup \Sigma_{S_1} \cup \cdots \cup \Sigma_{S_n}$ is a tuple $\mathfrak{M} = \langle G, S, I, M \rangle$, where*

- *G: is a set of relational symbols in Σ_G called* global schema*.*
- *S: is a set of data sources $S = \{S_1, \ldots S_n\}$ based on vocabularies $\Sigma_{S_1}, \ldots, \Sigma_{S_n}$, respectively.*
- *I: is a set of global integrity constraints $I = \{\mathcal{I}_1, \ldots \mathcal{I}_n\}$ expressed in the language of Σ_G.*
- *M: is a set of* schema mappings *of the form*

  $$R \leftarrow \Psi,$$

  where Ψ is a first-order expression constructed with the symbols in $\Sigma_{S_1} \cup \cdots \cup \Sigma_{S_n}$, and R is a predicate from Σ_G.

[2] We will sometimes refer to it as mediator-based system.

The declarative reading of the rule $R \leftarrow \Psi$ is as follows: for all tuples $\bar{t}$ such that Ψ is true, then $R(\bar{t})$ is also true. Schema mapping rules specify the ontological relationship between global and local predicates. Since in GeneReg the global schema acts as a data target for the information in the sources, we specify the schema mappings following a GAV approach.

The semantics of a data Integration system $\mathfrak{M}$ is defined as the first-order theory $\mathcal{M} = S \cup I \cup M$. In general $\mathcal{M}$ is an incomplete theory, i.e., there exists a formula over Σ which is not nor its negation entailed by the theory $\mathcal{M}$. In the current context, this is a desired property: given the incomplete information available in the data sources, it is expected that some data will be missing at the global level.

3 GeneReg: The Application at Work

3.1 The Global Schema

In GenReg the global schema G is the result of modelling the regulation process in the *Saccharomyces cerevisiae* with an entity relationship diagram (see [16] for a complete description of the modelling). It consists of the following relations:

- $condition\_info/7$: This relation specifies the parameters under which the experiments in the different sources were carried out.
- $gene\_condition/6$: Establishes the relationship between a gene, an expression and the conditions under which the relationship is measured.
- $gene\_regulator/3$: Associates a probability to a gene-regulator pair.
- $gene\_motif/3$: Provides a score for the likelihood of presence of a motif in the promoter of a gene.
- $gene\_expression/4$: Refers to the Loess normalized logarithm [24] of the ratio of the expression values in the red channel versus the green channel.
- $yeast\_genome/2$: Represents the genes of *Saccharomyces cerevisiae*.

3.2 Data Sources

3.2.1 Microarray data

Microarrays measure a snapshot of the activity of a cell by measuring the abundance of each type of mRNA molecule (which also gives an indirect and imperfect picture of the protein activity). In the past few years, microarray technology has emerged as an effective technique to measure the level of expression of thousands of genes in a single experiment. Because of their capacity to monitor many genes, microarrays are becoming the workhorse of molecular biologists studying gene regulation. One popular microarray technology platform, cDNA microarrays, uses two samples: a reference and a test sample (e.g., normal versus malignant tissue). These samples are labelled using distinct fluorescent molecules (green and red) and hybridized to the microarray. Relative amounts of a particular gene transcript in the two samples are determined by measuring the signal intensities detected at both fluorescence wavelengths and calculating the ratios. Microarray data is extracted from the Stanford Microarray repository [19]. The database is public and a mirror of the data is available in MS Excel format files.

3.2.2 Motif data

Motifs are short, conserved DNA-sequences. They are recognised by the regulators and therefore are very important to help unravelling the transcriptional regulatory network. The presence of a motif in the promoter region of a gene is indicated by means of a score or p-value. Motif data are retrieved from Kellis et al. [10] and stored in an internal repository at the KU Leuven after processing as described in Lemmens et al. [12].

3.2.3 ChIP-chip data

Transcription factors are proteins that bind to regulatory motifs in the promoter region of a gene to modify the rate of transcription of a gene. Chromatin immunoprecipitation on chip (or ChIP-chip) provides information on the direct physical interaction between a regulator and the promoter regions of its target genes. As such, ChIP-chip data indicate which proteins could be the regulators of a particular gene. The ChIP-chip data represent the probability (p-value) that a particular regulator binds to the promoter region of a particular gene. It is available at jura.wi.mit.edu/young_public/regulatory_code/GWLD.html and the format of the data is available in MS Excel, Matlab or plain text.

3.3 Schema Mappings and Integrity Constraints

Schema mappings describe the relationships between the global ontology and the local sources. To facilitate the data exchange between the ontologies, in GeneReg we have adopted the GAV paradigm –the global schema acts as a target for the data in the sources. As argued in [22], GAV is also the preferred option in settings where the number of sources remains constant, as it is the case in GeneReg. The mappings are defined below. Global primary keys [1] are underlined on the relevant variables.

$$\begin{aligned}
condition\_info(\underline{x_1}, x_2, x_3, x_4, x_5, x_6, x_7) &\leftarrow microarray(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}).\\
gene\_condition(\underline{x_8, x_1, x_9, x_{10}, x_{11}, x_{12}}) &\leftarrow microarray(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}).\\
gene\_regulator(\underline{x_8, x_1}, x_{13}) &\leftarrow chIP - chip(x_8, x_{14}, x_1, x_{13}).\\
gene\_motif(\underline{x_8, x_{15}}, x_{16}) &\leftarrow motif - locator(x_{17}, x_8, x_{15}, x_{16}).\\
gene\_expression(\underline{x_8, x_1, x_{18}, x_{19}}) &\leftarrow rFiles(x_8, x_1, x_{18}, x_{19}) \wedge genelist(x_8, x_{20}).\\
yeast\_genome(\underline{x_8}, x_{20}) &\leftarrow genelist(x_8, x_{20}).
\end{aligned}$$

Where the variables $x_1, \ldots x_{20}$ represent[3]:

Variable: GeneReg's name	Description	Variable: GeneReg's Name	Description
x_1 : Condition ID	Experiment's ID	x_{11} : Ch1Background	Channel1's background
x_2 : Exp Name	Experiment's name	x_{12} : Ch2Background	Channel2's background
x_3 : Category	Experiment's classif	x_{13} : Probability	Probability of occurence
x_4 : Subcategory	Experiment'subclassif	x_{14} : IntergenicNum	Intergenic number
x_5 : ExpDescription	Experiment's conditions	x_{15} : MotifID	Motif's ID
x_6 : Ch1Description	Channel1's content	x_{16} : Score	Motif in promoter
x_7 : Ch2Descritprion	Channel2's content	x_{17} :Identification	Source's key
x_8 : GeneID	Gene's ID	x_{18} : LogNor_A	Log ratio Ch1/Ch2
x_9 : Ch1Intensity	Channel1's intensity	$x19$: LogNor_M	Logmean ratio Ch1/Ch2
x_{10} : Ch2Intensity	Channel2's intensity	x_{20} : GeneName	Gene's name

3.4 System

GeneReg [5] has been developed for automatic management of data obtained from high-throughput techniques related to the gene regulation process in the organism *Saccharomyces cerevisiae*. The requirements comprise the data retrieval from the data-source, the pre-processing of information by means of wrappers, the global database population and finally the querying process. We describe each task separately.

- *Data retrieval.* Once all data sources are located and their availability verified by the application, the downloading of the relevant data files starts. The process is executed off-line and is performed every time new data is available in the sources. The information is then stored in local files for the preprocessing by the wrappers.

- *Data processing or wrapping.* A wrapper is a module that receives input data in a given format and translates it into a different data model (see [11] for a thorough discussion on wrappers). Its purpose is to present data in a suitable format for further processing. GeneReg's wrappers present the data-stream extracted from the data-sources as single relations of a database. In this way the population of the global schema can be directly executed by means of the schema mapping. In GeneReg, there exists one wrapper for each data source.

[3] For a full description of each attribute in the sources and the global schema see [16].

- *Data storage*. Once the data is extracted from the data sources and stored in a suitable format by GeneReg's wrappers, the information is migrated to the global database. To perform its task, the process of populating the global schema takes the wrapped data and exploits the ontological knowledge encoded in the rules of the schema mapping.

- *Querying Process*. With the data stored in the global database, the system is ready to received queries (by humans or automated reasoners). Since GeneReg relies in standard relational databases technology, the query processing can be performed by well-understood languages as SQL or Datalog.

4 Platform

GeneReg was developed for the Windows platform using the Java development kit (6.0). Two external libraries extend the basic functionality: the package POI JAKARTA for processing Excel files from the Stanford repository; and SFTP, for the management of ftp accounts. The database management system relies on MySQL server 5.0. The communication between the application and the database server is realized by MySQL/Java connector (5.0.3).

4.1 Execution and Performance

GeneReg downloads the information from the data sources at user's request. Not surprising, the total time required to complete a full upload of data will depend on the quality of the Internet networks and the load of the data sources servers. In average, the full processing of the motif database takes twenty minutes. Experiments from the microarray download at a rate of one per hour. The Excel files from the Stanford database are retrieved at a rate of four per minute. The combined time required for data processing by wrappers and the population of the global schema is neglectable in comparison with the time required for a full upload.

4.2 Soundness and Completeness

The soundness of GeneReg is measured with respect to the correctness of the data available in the global database. Since the data is extracted from the sources by means of *semantic* information (the GAV schema mapping), the only potential source of incorrectness in GeneReg are the data sources themselves. The investigation of the level of soundness of the local sources is beyond the scope of this work.

The global database produced by GeneReg is incomplete. This follows directly from the definition of mediator-based systems adopted in section 2.2. In practical terms, this means that there will be tuples that are true in the real world but are not present in the global database. This is not really surprising since one would expect that the data sources from which the information for the global database is extracted are themselves incomplete. Enormous effort has been put by the databases community to address the general problem of incompleteness. The interested reader is referred to [6] for a good survey on this topic.

5 Related Work

Heterogeneous biological data is distributed over different web sites and databases, making it very hard for the biologist to get access to all these publicly available data simultaneously. Therefore data integration of these heterogeneous data is of great importance to the biologist.

Although generic tools as INFOMIX [4] and Information Manifold [14] provide a friendly and semantically well-founded environment to integrate data, with these tools the process of adding a new data source ultimately relies on the user/programmer -for instance, to define the data source's wrapper. Given the highly heterogeneous and unstructured nature of the data sources required by users of GeneReg, the use of a generic tool would require not only the definition of new wrappers from scratch, but also their integration with an already existing application, with all the technical costs that such procedures would involve.

In the biological domain, data integration approaches such as BioGuideSRS [3] and OmicBrowse [21] have been developed to assist scientists with their search for relevant data within external sources (for example linking a gene to a disease). With these tools, the user can easily obtain an integrated view of distributed data servers. Although these kinds of platforms are important and useful, they usually integrate information about one gene, including the function of the gene product, links with diseases, homology to other organisms, genomic position etc. However, they do not allow a biologist to extract, for example, all expression levels for all genes of a particular organism in all conditions ever tested (i.e. all available microarrays for that organism). The BASE tool [17] provides an integrated framework for storing and analysing microarray information, but this tool also doesn't automatically collects all data. Publicly available databases that store microarray data in larger amounts do exist (for instance, the Stanford Microarray Database), but even from these databases it is hard for a biologist to download each single microarray (one by one) or to keep track of new experiments. Moreover, the data is again distributed over several such databases with overlapping content. Therefore, GeneReg was designed to automate the data collection process for several types of high throughput data sources.

Access to these high-throughput data is of utmost importance to bioinformaticians. Recently, several network inference methods and module detection tools have been developed that all use microarray data to reconstruct the regulatory network. Some of them combine microarray data with other available high-throughput data like ChIP-chip or motif data. Because creating and maintaining compendia of all available data is such a difficult and time-consuming task, most network reconstruction algorithms are currently being benchmarked (and applied) on subsets of the data only. However, it is crucial to incorporate as much data as possible to obtain more insight into the properties of different network inference methods and, more importantly, into the nature of the transcriptional network itself.

6 Discussion and Future Work

We have presented an implementation of a mediator-based system for integrating data sources related to the DNA transcription process in *Saccharomyces cerevisiae*. Our system follows the GAV paradigm to link the global ontology and the data sources. The information is instantiated in the global schema, which can be used by bioinformaticians for data mining purposes. In its current state GeneReg is only in a first development stage aiming at a framework for regulation process analysis in several organisms and data management oriented to a shared source of molecular knowledge. Future improvements of GeneReg include:

- GeneReg now gathers information for yeast only, but of course collecting data for more organisms is one of the possible and interesting future challenges. This extension not only would involve the addition of new data sources but the complete redefinition of the global schema.
- The data in the sources is extracted from high-output techniques and present a high level of uncertainty. In order to reduce this uncertainty, statistical processing of data seems to be an interesting approach. Specifically, the definition of regulatory modules created in the statistical process could serve as benchmark.
- At the technical level, an incremental extraction of data from the sources is highly desirable property. Since GeneReg downloads all the data from scratch, the populating phase is time consuming. Incremental extraction would improve performance considerably by updating only the relevant tuples. Another technical improvement relates to the possibility of updating the description of old sources or the addition of new ones.

The current state and future line of work of GeneReg should eventually converge to an intelligent and distributed system capable of inferring new knowledge over data statistically processed, based on a sound and well-founded semantic framework.

References

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley Publishing Company, 1995.

[2] D. Cavalieri and C. De Filippo. Bioinformatic methods forintegrating whole-genome expression results into cellular networks. *Drug Discovery Today*, 10:727–734, 2005.

[3] S. Cohen-Boulakia, O. Biton, S. Davidson, and C. Froidevaux. Bioguidesrs: querying multiple sources with a user-centric perspective. *Bioinformatics*, 23:1301–1303, 2007.

[4] Nicola Leone et al. The infomix system for advanced integration of incomplete and inconsistent data. In *SIGMOD Conference*, pages 915–917, 2005.

[5] GeneReg. Obtainable via www.cs.kuleuven.ac.be/˜alvaro/Gene_Regulation2.zip.

[6] Gösta Grahne. Information integration and incomplete information. *IEEE Data Eng. Bull.*, 25(3):46–52, 2002.

[7] D. Greenbaum, N.M. Luscombe, R. Jansen, J. Qian, and M. Gerstein. Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Research*, 11:1463–1468, 2001.

[8] M.J. Herrgard, M.W. Covert, and B.O. Palsson. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Research*, 13(11):2423–2434, 2003.

[9] Z. Joseph, G.K. Gerber, T.I. Lee, N.J. Rinaldi, J.Y. Yoo, F. Robert, D.B. Gordon, E. Fraenkel, T.S. Jaakkola, R.A. Young, and D.K. Gifford. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 21:1337–1342, 2003.

[10] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and Lander E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423:241–254, 2003.

[11] Nicholas Kushmerick. Wrapper induction: Efficiency and expressiveness. *Artif. Intell.*, 118(1-2):15–68, 2000.

[12] K. Lemmens, T. Dhollander, T. De Bie, P. Monsieurs, K. Engelen, B. Smets, J. Winderickx, B. De Moor, and K. Marchal. Inferring transcriptional module networks from chip-chip-, motif- and microarray data. *Genome Biology*, 7(5):R37, 2006.

[13] M. Lenzerini. Data integration: A theoretical perspective. In *21st PODS, June 3-5, Madison, Wisconsin, USA*, pages 233–246, 2002.

[14] A. Levy, A. Rajaraman, and Ordille J.J. Querying heterogeneous information sources using source descriptions. In *VLDB-96*, volume 1, pages 251–262, 1996.

[15] H. Lodish, D. Baltimore, A. Berk, S. Zipursky, P. Matsudaira, and J. Darnell. *Molecular cell biology*. Scientific American Books, New York, USA, 1995.

[16] D. Pastor. Integration of Heterogeneous Data Sources of a DNA Transcription Process. Master thesis, K.U.Leuven. Obtainable via www.cs.kuleuven.ac.be/˜alvaro/thesis_pastor.zip, 2007.

[17] L.H. Saal, C. Troein, J. Vallon-Christersson, S. Gruvberger, A. Borg, and C. Peterson. Bioarray software environment (base): a platform for comprehensive management and analysis of microarray data. *Genome Biology*, 3:software0003.1–0003.6, 2002.

[18] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34:166–176, 2003.

[19] Stanford MicroArray Database. Obtainable via genome-www5.stanford.edu/.

[20] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences of the USA*, 101:2981–2986, 2004.

[21] T. Toyoda, Y. Mochizuki, K. Player, N. Heida, N. Kobayashhi, and Y. Sakaki. Omicbrowse: a browser of multidimensional omics annotations. *Bioinformatics*, 23:524–526, 2007.

[22] J. Ullman. Information integration using logical views. *Theoretical Computer Science*, 239(2):189–210, 2000.

[23] T. Van den Bulcke, Lemmens K., Y. Van de Peer, and K. Marchal. Inferring transcriptional networks by mining 'omics' data. *Current Bioinformatics*, 1(3):301–313, 2006.

[24] Y.H. Yang et al. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30:e15, 2002.