BMC Bioinformatics

Proceedings

BioCAD: an information fusion platform for bio-network inference and analysis

Doheon Lee*, Sangwoo Kim and Younghoon Kim

Address: Department of Bio and Brain Engineering, KAIST, 373-1 Guseong-Dong, Yuseong-Gu, Daejeon, 305-701, Republic of Korea

Email: Doheon Lee* - dhlee@biosoft.kaist.ac.kr; Sangwoo Kim - swkim@biosoft.kaist.ac.kr; Younghoon Kim - yhkim@biosoft.kaist.ac.kr

* Corresponding author

from First International Workshop on Text Mining in Bioinformatics (TMBio) 2006
Arlington, VA, USA. 10 November 2006

Published: 27 November 2007

BMC Bioinformatics 2007, **8**(Suppl 9):S2 doi:10.1186/1471-2105-8-S9-S2

This article is available from: http://www.biomedcentral.com/1471-2105/8/S9/S2

Abstract

Background: As systems biology has begun to draw growing attention, bio-network inference and analysis have become more and more important. Though there have been many efforts for bio-network inference, they are still far from practical applications due to too many false inferences and lack of comprehensible interpretation in the biological viewpoints. In order for applying to real problems, they should provide effective inference, reliable validation, rational elucidation, and sufficient extensibility to incorporate various relevant information sources.

Results: We have been developing an information fusion software platform called BioCAD. It is utilizing both of local and global optimization for bio-network inference, text mining techniques for network validation and annotation, and Web services-based workflow techniques. In addition, it includes an effective technique to elucidate network edges by integrating various information sources. This paper presents the architecture of BioCAD and essential modules for bio-network inference and analysis.

Conclusion: BioCAD provides a convenient infrastructure for network inference and network analysis. It automates series of users' processes by providing data preprocessing tools for various formats of data. It also helps inferring more accurate and reliable bio-networks by providing network inference tools which utilize information from distinct sources. And it can be used to analyze and validate the inferred bio-networks using information fusion tools.

Background

Understanding internal networks of a given system is one of the ultimate goals in biological studies. Inferring precise networks includes both processes of assigning functional annotations to each element of networks and predicting flows of causal effects between those elements. To complete these processes, plenty of data from various sources and proper algorithms for network inference are needed.

Most of the studies of inferring biological networks have taken computational and statistical approaches. In the case of inferring genetic regulatory networks, microarray expression profile data has been widely used to look into the internal activities of cells, and a lot of studies have

been done to apply computational algorithms for network inference to such data. So far, Bayesian network has been widely used because it has sound mathematical basis and the characteristic of noise resistance [1,2]. Other computational techniques like correlation metric construction [3], dynamic Bayesian network [4,5], S-system [6,7], Boolean network [8], logic gate model [9] and Petri-net [10] were also applied to inferring or modeling genetic regulatory networks from microarray gene expression data. Besides, literature information also has been used to build biological networks [11,12].

Although such techniques for inferring biological networks have been developed and improved up to now, several problems still exist. First, those inferred networks usually contain many false inferences and they are mainly due to the lack of information (the amount of available data is very limited in general). Most of available microarray data does not contain enough number of experiments to infer reliable networks when considering the large number of genes. Noise problems in preprocessing and information loss in inference processes are also reasons of such false inferences. Second, the relationships such as dependency, coherence or causality in the inferred networks can be ambiguous. The network itself usually does not elucidate why those edges exist; how strongly the elements affect the others; and which of activation or repression they indicate.

Because the network inference from single data source has such limitations mentioned earlier, there have been several studies of utilizing additional information. Hartemink et al [13] used location and expression data together to infer genetic regulatory networks. Kato et al [14] proposed a kernel-based method for supervised network inference based on multiple types of biological datasets such as gene expression, phylogenetic profiles and amino acid sequences. Xing et al [15] also used gene expression and sequence data to infer gene regulatory networks.

Information fusion processes can be also used for further analysis of inferred networks after the inference process. Validating inferred networks requires additional information sources such as annotation database, literature and other already known networks. Text mining tools play an important role in utilizing such information sources. Analyzing inferred networks reveals the characteristics of networks such as connectivity, topology, network motifs and dynamics. Using network validation and analysis processes enables the inferred networks to be more accurate, reliable and rationally elucidated.

However, the information fusion process is not always easy to be applied in general. First, the format of available data is not unified. For example, there are more than six data formats which are used widely for microarray expression profile data including SOFT format of NCBI GEO database [16], Mage-ML [17], GenePix format, Spot format, conventional tab delimited or comma separated format. This variety of data format becomes more serious when we consider data-to-data conversion in the data preprocessing and network inference processes. Second, we need to have various algorithms and tools to deal with the diverse types of data including microarray expressions, mass spectrometry, and literature information. Thus it is not easy to find optimal tools for network inference and validation with respect to the various data formats and characteristics.

About these problems, several works have been proposed to serve integration platforms where different types of data and processing algorithms are used. Cytoscape [18] is a plug-in oriented information fusion platform. Its core function is network visualization, but a set of plug-ins enables one to assay microarray data and annotate inferred networks. Systems Biology Workbench (SBW) [19] also tries to connect various tools for given data. The approach of SBW is to connect programs each other tightly with a common data model, which is SBML. Taverna Project [20] has a little bit different characteristic, which serves workflows defined by Web Services technologies. Taverna enables users to define their own biological workflows, connect to the designated Web Services so that a series of processes can be done in one phase. Although previous information fusion platforms were successful in some aspects, several important features have to be considered for the network inference and analysis processes in information fusion platforms. An information fusion platform should provide effective modules that users can easily use for reliable inference, validation and elucidation of bio-networks. Further, sufficient extensibility and well defined workflows are also required to help users incorporate various information sources. Cytoscape and SBW provide good network inference and analysis tools via TCP/IP socket connection and in the form of plug-in modules. However, both platforms do not provide the workflow feature and sufficient extensibility such as Web Services in Taverna. Taverna has very good extensibility with user definable workflows. But its target is too general so that users cannot easily apply it to network inference and analysis. In this study, we propose an information fusion platform named BioCAD, which supports the whole processes of network inference and analysis with good extensibility and the workflow features.

Results and discussion

BioCAD system architecture

BioCAD is an integration environment where various functional modules are inter-operated. The name of Bio-

CAD represents Bio Computer Aided Design because the processes of network inference and analysis are a part of general designing processes. BioCAD has been developed on the Netbeans platform using the Java language. Basically, all the BioCAD modules are independent of each other except the structural modules used for software building and user interface. This feature of modularization enables the BioCAD software to have good extensibility. The aspect of BioCAD user interface is shown in Figure 1.

BioCAD functional modules are divided into three major categories – Data preprocessing module, network inference module, network analysis module. Data preprocessing module takes charge of modification of given data to the best form for subsequent works. This includes data-filtering, re-scaling, taking logs and normalization with respect to given data formats. PCA analysis and general clustering and classification tools can be used in users' needs. Network inference module has a set of tools for inferring network shaped structures from other types of data.

Currently, inference tools which implemented Temporal Association Rule Mining [21] and MONET [22] are supported. Other inference tools are being developed and planned to corporate via Web Services. Network analysis module includes validating inferred network using external information such as protein-protein interaction and text-mining data. And static/dynamic network analysis

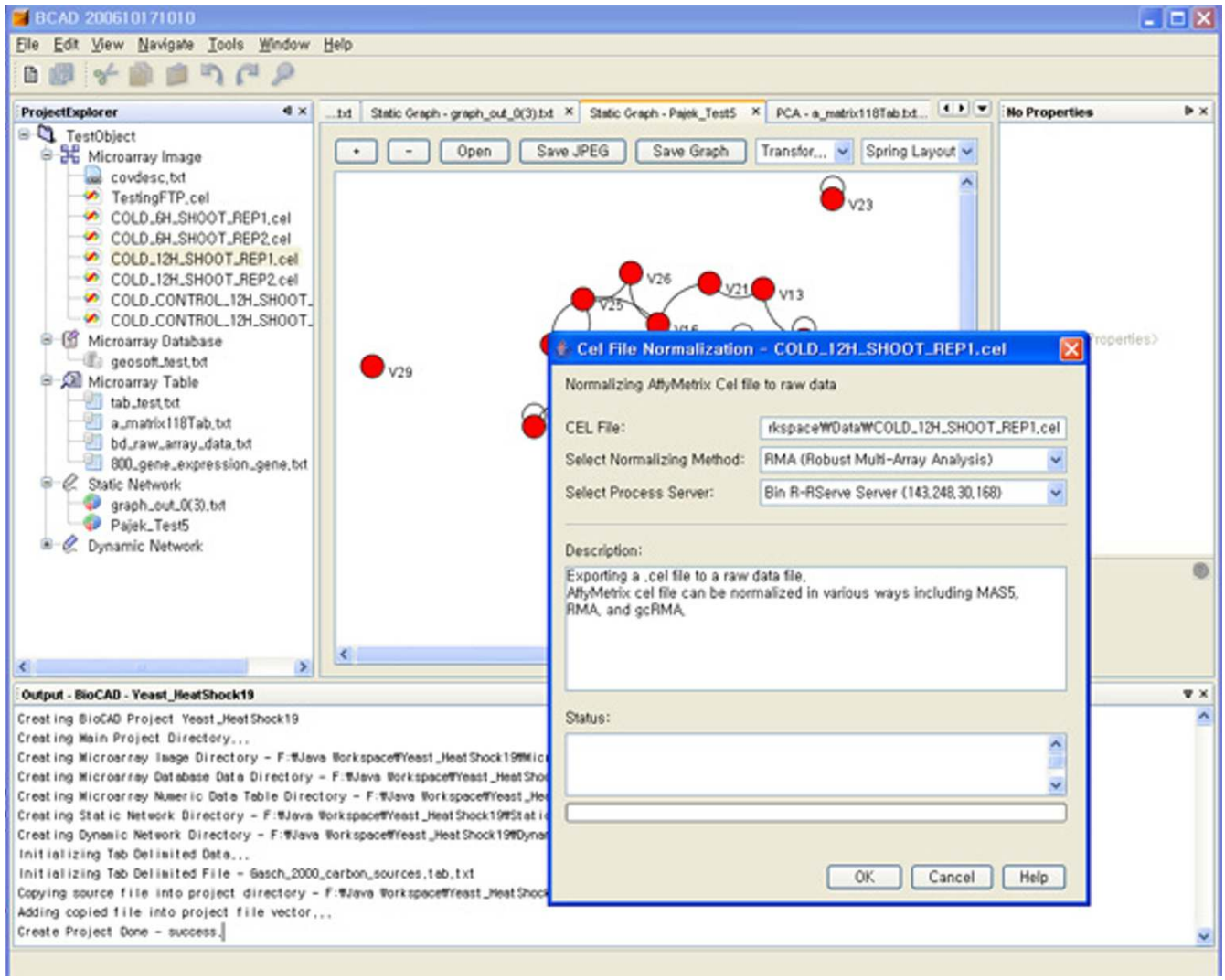

Figure 1
User Interface of BioCAD. BioCAD is composed of various data preprocessing and functional modules. Most of jobs are selected and executed in the Project Explorer Window (left)

algorithms such as network motif analysis, network characteristic analysis and network dynamics analysis are planned to be implemented. Using these modules, users can advance their own jobs with their own data along with the predefined workflows.

Inferring networks using network inference tools

BioCAD provides whole processes used for network inference. A series of processes provided by BioCAD software is called a BioCAD project. A BioCAD project consists of source data and processed data from working modules. When users have their own data sources and want to infer a network from the source, BioCAD project is created by a new project wizard. Using the wizard, the user can assign the data source from various data formats the BioCAD software provides, and start the user's job along with the workflow. Data formats supported in BioCAD projects is shown in Table 1.

As we mentioned earlier, there have been a lot of network inference studies, which can be included in BioCAD as network inference modules. Boolean networks map the activity level of a gene into a binary state, on or off. Although the constructed Boolean network can simulate the flow of regulations, the binary representation of state and synchronous transition is two major drawbacks. Other algebraic approaches including differential equation model, S-system can construct very accurate networks which are able to be simulated. However, most of data have too few samples compared to the number of genes, it is meaningless to extract that much of information from the data, and the inference process is to be too time-consuming.

As importing network inference modules into BioCAD, we considered two major features – usability and accuracy. For the usability's sake, algorithms which need too much time to calculate such as S-system, differential equation model, conventional Bayesian network model are excluded. And for accuracy's sake, Boolean network is excluded due to the limitation of network notation. Currently, we focus on ARACNE and MONET. ARACNE [23] is a novel algorithm using microarray expression profiles and mutual information processes between a pair of random variables. ARACNE algorithm shows good performance compared to the algorithm complexity and the result represents sufficient information of causes and affections. MONET is basically a Bayesian Network algorithm. However, MONET has adopted a divide-and-conquer approach to alleviate the dimensionality problems. MONET shows good usability due to its modularizing processes and noticeable improvement of accuracy.

Assuming that a user wants to infer a network starting from NCBI's GEO SOFT file, the user connects to a Web Services tool that imports the file from the Web. From the microarray database file, microarray expression profile can be extracted. Next, the user can preprocess the extracted profile data using data preprocessing modules either provided in BioCAD's built-in tools or supported Web Services tools. BioCAD provides effective preprocess tools associated with the BioConductor package. Finally MONET starts inferring process with the user's request. MONET uses Gene Ontology database in its inferring process. Because BioCAD does not involve the MONET module in the form of built-in tool, MONET's information fusion process with GO term can be accomplished in the specified MONET server. The inferred Bayesian network is shown both in graph and table view. This network data is also a part of BioCAD project, and can be used for subsequent processes.

Analyzing networks using information fusion tools

In the BioCAD project, inferred bio-network is treated as a new source for subsequent analysis and validation processes. One good validating method is inspecting network's relations utilizing text mining tools. There have been various studies in applying text mining techniques to the bioinformatics area by means of information extraction, information retrieval and natural language processing (NLP). Donaldson et al [24] used a support vector machine to extract protein-protein interaction data. Saric et al [12] created rule based system STRING-IE to construct gene and protein regulatory networks from Medline database. The text mining techniques are also used for extracting gene/protein's information and automatic annotation [25].

Table 1: Supported Data Formats in BioCAD Project

Data Category	Data Format
Microarray Image File	TIFF, AffyMetrix Cel File
Microarray Database File	SOFT
Microarray Expression Profile	Tab Delimited, CSV
Static Network	SBML, GML, Pajek
Dynamic Network	SBML

BioCAD Supports reading and writing most commonly used file formats. Some of the file formats and additional formats are being implemented.

Currently BioCAD is equipped with a text mining tool which finds regulation or interaction information between two genes from literature search. The constructed network in the previous step is to be examined through the validating step by putting a pair of genes which are connected in the network into the text mining tool. As a result, we can find out whether each network connection has its supporting literature information and in what kind of relation it is connected.

Information fusion can be used between different types of large-scale data. Cohen et al [26] used chromosome correlation maps to express patterns of genes of the same chromosome. Drawid et al [27] used protein subcellular localization data to inspect the relationship with gene expression profiles. Lotem et al [28] integrated protein-protein interaction and transcription regulation data of *S. cerevisiae* to find specific regulatory relations, such as positive and negative feedback circuits. Using those inter-data analysis models, separated networks or databases can be integrated to elucidate more specific and accurate relations.

The network analysis tool provided in BioCAD is integration a genetic regulatory network with its corresponding protein-protein interaction map, named Bio-viaduct. Bio-viaduct defines a pathway where a gene can affect another gene via transcriptional regulation and protein-protein interactions. For example, when there is a directed edge from gene A to gene B in the inferred network, it searches paths from expressed protein of gene A to gene B's transcription factor connected by intermediate protein(s), Bio-viaduct module is also provided via Web Services so that the user can proceed only by operating a command, invoking the remote Bio-viaduct server to receive the source network and compute the pathways using the server-side protein-protein interaction information

Extending modules with Web services and BPEL workflows

One of the most potential ability of BioCAD is the good extensibility from applying the Web Services technologies. Web Services is a software system designed to support interoperable machine-to-machine interaction over a network. Because this definition encompasses many different systems, in common usage the term usually refers to those services that use SOAP-formatted XML envelopes and have their interfaces described by WSDL. Even though Web Services is another attempt to standardize the Remote procedure call protocol (RPC) between platforms by piggybacking on the near-universally deployed HTTP protocol, it has its own advantages; it is loosely coupled thereby facilitating a distributed approach to application integration and it is Independent of the client side technologies used.

When a new network inference or analysis module is required, BioCAD can register the target tools using the module's WSDL file. Every public Web Services program has its WSDL file to describe the program's functionalities and a required set of input. In the case of that a target tool is not in the form of public Web Services, BioCAD can read the target program's compiled file and create the WSDL file. Currently, several modules are in the process of integration to BioCAD. This extensibility enables the BioCAD to keep up with the new technologies of data preprocessing, network inference and network analysis.

BioCAD uses BPEL (Business Process Execution Language) [29] as the description language of workflows. BPEL is standard language that defines business process and execution in Web Services environment. BPEL provides a rich vocabulary for defining processes and has several features which are not found in programming languages. Also, the Netbeans platform supports designing BPEL processes since version 5.5. A sample BPEL flow defined in BioCAD is shown in Figure 2. BioCAD provides a Project Flow Window which makes the users can monitor the current status and available processes in the workflow.

Conclusion

We have proposed an information fusion platform named BioCAD. It provides a convenient infrastructure for network inference and network analysis. We showed three major profits that can be obtained from using BioCAD. First, it automates series of users' processes by providing data preprocessing tools for various formats of data. The RCP based user interface and workflows make it easier and more familiar to use the software. Second, BioCAD helps inferring more accurate and reliable bio-networks with providing network inference tools which utilize information from distinct sources. We showed a process of Bayesian network construction from an entry of microarray database using MONET which makes use of gene annotation information. Third, BioCAD can be used to analyze and validate the inferred bio-networks. Text mining and Bio-viaduct tools are in capable of integrating different types of information into the constructed networks.

One of the most potential features of BioCAD is its extensibility. Because the whole functionalities of BioCAD are modularized, any other tools which provide related functions such as network inference and network analysis and other types of functions including network visualization and network topology analysis can be easily added. Due to the workflow facility, those all new modules also can be integrated to the currently provided modules.

Competing interests

The authors declare that they have no competing interests.

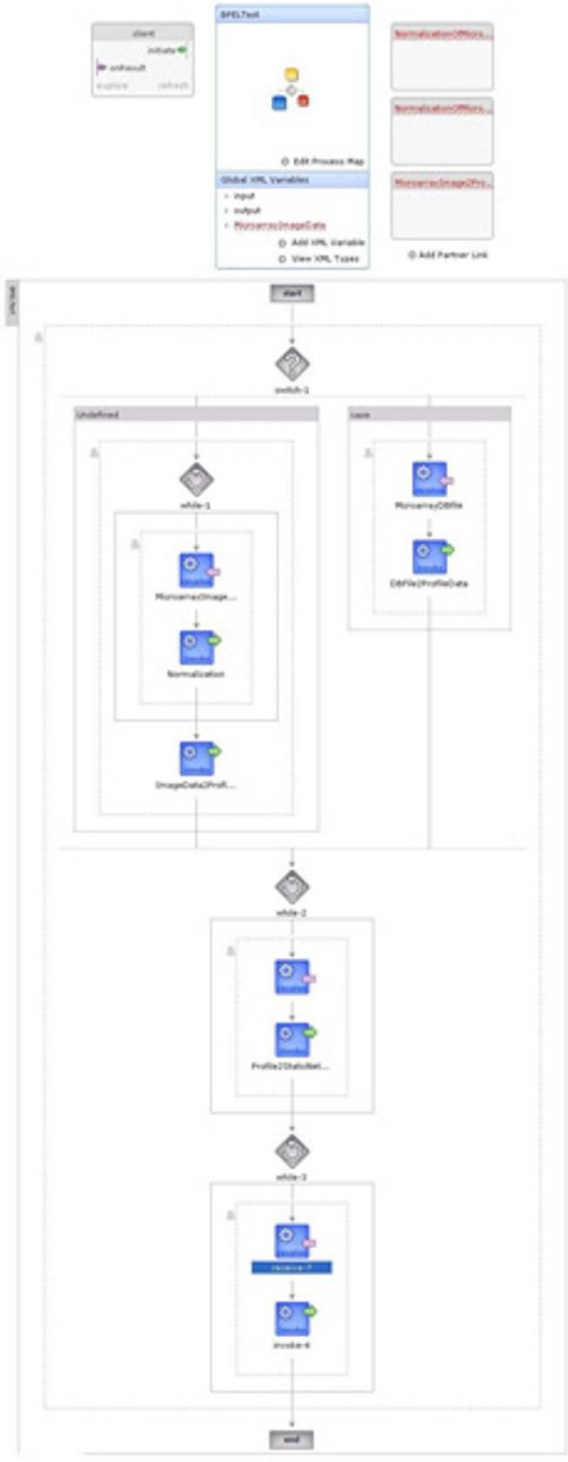

Figure 2
BPEL Workflow of BioCAD. A simplified diagram of bio-reverse engineering processes of BioCAD. Further workflows are defined while users add, remove or repeat each of the jobs.

Authors' contributions

DL developed the main idea of component based framework for reverse engineering. SK implemented core and functional modules of the program. YK implemented network inference modules that this program makes use of.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation(KOSEF) through the National Research Lab. Program funded by the Ministry of Science and Technology (No. 2005-01450). We would also like to thank CHUNG MoonSoul Center for BioInformation and BioElectronics for providing research and computing facilities.

This article has been published as part of *BMC Bioinformatics* Volume 8 Supplement 9, 2007: First International Workshop on Text Mining in Bioinformatics (TMBio) 2006. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/8?issue=S9.

References

1. Friedman N, *et al.*: **Using Bayesian Networks to Analyze Expression Data.** *Journal of Computational Biology* 2000, **7(3–4):**601-620.
2. Pena JM, Bjorkegren J, Tegner J: **Growing Bayesian network model of gene networks from seed genes.** *Bioinformatics* 2005, **21(Suppl 2):**ii224-ii229.
3. Arkin A: **A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements.** *Science* 1997, **277(5330):**1275-1279.
4. Madigan D, Raftery AE: **Model Selection and Accounting for Model Uncertainly in Graphical Models Using Occam's Window.** *Journal of the American Statistical Association* 1994, **89(428):**.
5. Kim SY, Imoto S, Miyano S: **Dynamic Bayesian Network and Nonparametric Regression Model for Inferring Gene Networks.** *Genome Informatics* 2002, **13:**371-372.
6. Kikuchi S, *et al.*: **Dynamic modeling of genetic networks using genetic algorithm and S-system.** *Bioinformatics* 2003, **19(5):**643-650.
7. Kimura S, Hatakeyama M, Konagaya A: **Inference of S-system models of genetic networks from noisy time-series data.** *Chem-Bio Informatics Journal* 2004, **4(1):**1-14.
8. Lädesmäi H, Shmulevich I, Yli-Harja O: **On Learning Gene Regulatory Networks Under the Boolean Network Model.** *Machine Learning* 2003, **52(1):**147-167.
9. Bulashevska S, Eils R: **Inferring genetic regulatory logic from expression data.** *Bioinformatics* 2005, **21(11):**2706-2713.
10. Mayo M: **Learning Petri net models of non-linear gene interactions.** *Biosystems* 2005, **82(1):**74-82.
11. Saric J: **Large-Scale Extraction of Gene Regulation for Model Organisms in an Ontological Context.** In *Silico Biology* 2005, **5(1):**21-32.
12. Saric J, *et al.*: **Extraction of regulatory gene/protein networks from Medline.** *Bioinformatics* 2006, **22(6):**645.
13. Hartemink AJ, *et al.*: **Combining location and expression data for principled discovery of genetic regulatory network models.** *Pac Symp Biocomput* 2002, **7:**437-449.
14. Kato T, Tsuda K, Asai K: **Selective integration of multiple biological data for supervised network inference.** *Bioinformatics* 2005, **21(10):**2488-2495.
15. Xing B, van der Laan MJ: **A Statistical Method for Constructing Transcriptional Regulatory Networks Using Gene Expression and Sequence Data.** *Journal of Computational Biology* 2005, **12(2):**229-246.
16. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Research* 2002, **30(1):**207-210.
17. Brazma A, *et al.*: **ArrayExpress – A public repository for microarray gene expression data at the EBI.** *Nucleic Acids Research* 2003, **31(1):**68-71.
18. Shannon P, *et al.*: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13(11):**2498-504.

19. Hucka M, *et al.*: **The ERATO Systems Biology Workbench: enabling interaction and exchange between software tools for computational biology.** *Pac Symp Biocomput* 2002, **1:**450-461.
20. Stevens RD, Robinson AJ, Goble CA: **myGrid: personalised bioinformatics on the information grid.** *Bioinformatics* 2003, **19(Suppl 1):**i302-i304.
21. **Temporal Association Rule Mining** [http://biosoft.kaist.ac.kr/~hjnam/TARM/TARM.html]
22. Lee PH, Lee D: **Modularized learning of genetic interaction networks from biological annotations and mRNA expression data.** *Bioinformatics* 2005, **21(11):**2739-2747.
23. Margolin AA, *et al.*: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.** *BMC Bioinformatics* 2006, **7(Suppl 1):**S7.
24. Donaldson I, *et al.*: **PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine.** *BMC Bioinformatics* 2003, **4:**11.
25. Tamames J: **Text detective: a rule-based system for gene annotation in biomedical texts.** *BMC Bioinformatics* 2005, **6(Suppl 1):**S10.
26. Cohen BA, *et al.*: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.** *Nat Genet* 2000, **26(2):**183-6.
27. Drawid A, Jansen R, Gerstein M: **Genome-wide analysis relating expression level with protein subcellular localization.** *Trends Genet* 2000, **16(10):**426-30.
28. Yeger-Lotem E, Margalit H: **Detection of regulatory circuits by integrating the cellular networks of protein-protein interactions and transcription regulation.** *Nucleic Acids Res* 2003, **31(20):**6053-61.
29. **Business Process Execution Language for Web Services (BPEL), Version 1.1** [http://www-128.ibm.com/developerworks/library/specification/ws-bpel/]