In Silico Biology 9, 0007 (2009); ©2009, Bioinformation Systems e.V.

Classification of information fusion methods in systems biology

Jane Synnergren*, Björn Olsson and Jonas Gamalielsson

Systems Biology Research Centre, School of Life Sciences, University of Skövde, SE-541 28 Skövde, Sweden

* Corresponding author

Email: jane.synnergren@his.se

Edited by E. Wingender; received August 04, 2008; revised February 21, 2009; accepted February 23, 2009; published April 15, 2009

Abstract

Biological systems are extremely complex and often involve thousands of interacting components. Despite all efforts, many complex biological systems are still poorly understood. However, over the past few years high-throughput technologies have generated large amounts of biological data, now requiring advanced bioinformatic algorithms for interpretation into valuable biological information. Due to these high-throughput technologies, the study of biological systems has evolved from focusing on single components (e.g. genes) to encompassing large sets of components (e.g. all genes in an entire genome), with the aim to elucidate their interdependences in various biological processes. In addition, there is also an increasing need for integrative analysis, where knowledge about the biological system is derived by data fusion, using heterogeneous data sets as input. We here review representative examples of bioinformatic methods for fusion-oriented interpretation of multiple heterogeneous biological data, and propose a classification into three categories of tasks that they address: data extraction, data integration and data fusion. The aim of this classification is to facilitate the exchange of methods between systems biology and other information fusion application areas.

Keywords: information fusion, data fusion, data integration, systems biology

Introduction

Systems biology aims to understand the behaviour and interaction between various components of the living cell, such as genes, proteins and metabolites. When studying a biological system one typically perturbs it and use different high-throughput assays to identify interacting elements and elucidate their relationships. These diverse data are then used to infer biological networks to increase the knowledge of the system's behaviour. Much effort has been made to understand how these components interact. However, we are still far from practical applications due to problems with many false inferences, and the lack of a comprehensive interpretation from the biological points of view [1, 2].

Over the past years, development of high-throughput technologies have resulted in a shift of focus from single components (reductionism) to large-scale analysis, e.g. from studying single genes to analysis of large sets of genes and their interdependences in biological processes [3]. Various technologies provide different types of insights about the system, and fusion of data from several sources is therefore of great importance in the efforts to understand a particular biological process. In addition to high-throughput data there is also valuable information from e.g. small-scale experiments, curated databases, and computational predictions. By integration and fusion of data the dimensionality can be reduced and heterogeneous data interpreted in such a way that we gain highly useful information about the system under study.

The motivation for applying Information Fusion (IF) to a combination of data sources is that the resulting output should provide more knowledge about the investigated system than what would have been obtained by analysing the individual data sources separately [4]. A major challenge in this process is the inherent complexity in fusion of heterogeneous data. Specific attention must be paid to the biological relations between different types of data. Moreover, high-throughput technologies suffer from inherently high false-positive rates and each technology generates a unique set of systematic biases [1]. Another problem is that there are usually no curated data sets from which one can estimate the integration parameters [1]. Yet another challenge is to integrate expert knowledge that biologists possess about the biological process under investigation and fully utilize that in the data integration process [5]. Furthermore, the exponential growth of many biological databases due to rapidly developing large-scale technologies requires efficient database search algorithms. To fully utilize the valuable resources it is of critical importance to achieve an efficient workflow for in silico studies of complex biological processes [6].

Information fusion is a rapidly developing research field, for which several definitions exist. Here we apply the one proposed in [7]: *"Information fusion is the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making."* Data Fusion (DF) is another (older) term that is also used almost interchangeably with IF. However, IF usually has a slightly wider meaning than DF, which typically deals with combining raw data. IF is commonly associated with military applications but methods and algorithms from IF can most likely be advantageous also for analogous problems in the biological domain. Conversely, it is also probable that many of the methods developed for fusion of biological data are general enough to be useful in other IF domains. We propose that it would be advantageous for several domains if we could exchange methods across different research areas. To facilitate this, methods need to be general in structure and described in a structured and standardized way using a widely known vocabulary. Therefore, a thorough survey of methods and algorithms that are available for integration or fusion of various types of biological data is needed. There is also a need to investigate the applicability of typical IF methods to the fusion of biological data for scenarios

biological data is needed. There is also a need to investigate the applicability of typical IF methods to the fusion of biological data for scenarios where there is currently a lack of solutions. Is it possible to make minor modifications of these methods to achieve a better adaptation to biological problems? Finally, there is a need to develop novel methods for fusion of biological data with characteristics that require a customized solution.

Classification categories

As illustrated in Fig. 1, we propose three different categories for classification of IF-related methods. The first category is *Data Extraction* (DE), representing methods for retrieval of various types of heterogeneous data in an automated way. There are often many features available in many different databases, and these are extracted in an automated way by DE methods. Examples of application areas include extraction of sequence data, structural data, tissue expression data, and information about protein families. Furthermore, a distinction is made between methods for *Data Integration* (DI) and *Data Fusion* (DF). We adopt the definition of data integration that is proposed in MSN Encarta [8]; "*Data integration: the integration of data and knowledge collected from disparate sources by different methods into a consistent, accurate, and useful whole.*" We here classify a method as a DI method if it uses data from several sources, but uses these data sources individually, e.g. by verifying or filtering results derived from one data source by using results from other data sources. Generally, these data are from different abstraction levels of the system or are generated by different experimental techniques. The aim of the DI is to combine data from different sources and provide the user with a unified view of these data, which helps the user to get an overview and facilitates the accessibility of the available data [9].

In addition to DI methods there are also DF methods that typically use several data sources in combination and where the resulting output from the methods is dependent on all the various input data in combination. Here we base our categorization on the definition of data fusion formulated by [10]: "*Data fusion deals with the synergistic combination of information made available by various knowledge sources such as sensors, in order to provide a better understanding of a given scene.*" The aim of DF is to achieve an improved model that agrees with all available data for the biological process or system under study [9]. As an example of a DF method that is used in systems biology, one can cluster genes based on their expression profiles while also incorporating into the clustering process data about binding sites or regulatory elements. Genes with similar expression profiles will then be clustered together only if they also share common regulatory elements. Such a method that uses two or more types of data in combination in the same procedure is here classified as a DF method. To emphasize the distinction between DI and DF methods, Fig. 1 shows how a model of the studied system can either be based on a combination of inferences which have been drawn from different data sources separately and then integrated in the model, and/or based on inferences drawn in a process where the different data sources have been used simultaneously during the inference process. The synergistic combination of information from several sources in data fusion means that the amount of information obtained by combining a set of sources is greater than the amount of information provided by the individual data sources alone [4]. This can formally be expressed as

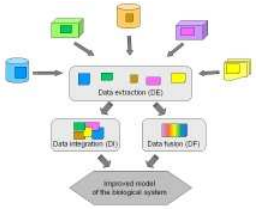
$$I\left(f(s_1,\ldots,s_n)\right) > \sum_{j=1}^{n} I(s_j)$$

where I is the amount of information, f is a function representing the fusion of n data sources and s_x is a specific data source. This can be contrasted to the situation in DI where there are no synergistic effects in terms of amount of information when several data sources are used, but where the integrated information still provides valuable information to the model representing the biological system (Fig. 1). This can formally be expressed as

$$I\left(i(s_1,\ldots,s_n)\right) = \sum_{j=1}^{n} I(s_j)$$

where i is a function representing the integration of n data sources.

Here we have surveyed a set of representative studies that propose methods for extraction, integration and fusion of heterogeneous data from various sources when analyzing biological systems. We have also classified these methods into three categories of tasks illustrated in Fig. 1, reflecting the types of problems addressed by the methods.



Click on the thumbnail to enlarge the picture

Figure 1: Schematic view of workflow: Different types of data are extracted from various data sources. These data are subsequently integrated or fused, either manually or automatically by various methods, with the aim to improve the model of the system under study. In DI methods, each data source is used separately to create or improve a model of the system under study, whereas DF methods use the different data sources simultaneously.

Information fusion methods

A rapidly increasing amount of functional genomic data is nowadays available in public databases and many important insights have been gained from analyzing them. Nevertheless, we are far from a full utilization of the great potential of these resources. The aim of a typical research project applying a systems biology approach can be to: (i) use two or more, often heterogeneous, data sets to try to reconstruct networks that describe the biological process; and (ii) try to understand the role of the genes and proteins participating in the process and how they interact to perform a specific function.

Solely retrieving various types of relevant available data for a large set of genes is in itself a demanding task and requires appropriate bioinformatics tools. Moreover, using data associated with an identified set of genes to derive an overall understanding of the underlying biological process by integrating or fusing various types of data constitutes a significant challenge. However, in recent years several tools have been

developed both for retrieving data from various sources and for integrating or fusing different types of functional genomics data in an automated or semi-automated manner [3]. One of the main problems with these tools is that they generate too many false inferences, and that there is a lack of methods for reliable validation in most application domains. Another problem is that the methods are often restricted to specific types of data or limited to a specific organism. Biological data are often heterogeneous in nature and characterized by inherent (and sometimes unknown) relationships, which complicates the information fusion process. Scalability is also an issue when designing the methods, since many biological databases grow exponentially. Much more effort must therefore be spent on developing efficient and automated methods for integration and fusion of various types of data, to allow researchers to better utilize the huge amount of valuable data that is being generated by the high-throughput technologies.

In the following sections we review a number of representative existing methods for extraction, integration and fusion of biological data. We also group them into three different categories of tasks, as illustrated in Fig. 1, corresponding to the types of problems the methods address.

Data extraction and integration methods

Retrieving relevant available data for subsequent data integration or fusion is a challenge even for a limited set of genes and it therefore needs to be efficiently facilitated. In addition, the need to integrate or fuse heterogeneous data from several sources into one common understanding of a biological system makes the task even more challenging. To address these demands, a number of data extraction and integration systems of various types have been developed that focus on different types of data extraction and integration. Few methods exist that perform DI in an automated manner. Commonly the integration is done manually or semi-automatically by the researcher, who also incorporates his or her professional knowledge in the DI process. Here we report on four methods from these categories which can be considered as representative examples of DE and DI methods. That the example methods are frequently used by the research community is indicated by their high citation indices, e.g. reference [3] and [11] have been cited by more than 100 other research papers. Additional examples of other methods from these categories are given in [12, 13, 14].

Data extraction methods

Zhang and co-workers [3] have developed a tool named WebGestalt (WEB-based GEne SeT AnaLysis Toolkit), which is an integrated data mining system for the management, information retrieval, organization, visualization and statistical analysis of large sets of genes. WebGestalt consists of four modules: (i) Gene set management, (ii) Information retrieval, (iii) Visualization, and (iv) Statistics. With the gene set management module, sets of interesting genes can be up-loaded, stored and retrieved again at a later stage. The information retrieval module can retrieve up to 20 different types of data for each gene in a gene set, including Gene Ontology (GO) annotations [15], tissue expression patterns, chromosomal distributions, different types of related pathways, relevant protein domains and references to related publications. The tool provides functionality for creation of subsets of genes from a gene set based on different criteria, such as GO categories, biochemical pathways or chromosome location ranges. The prime advantage of WebGestalt is the function to derive a tissue expression pattern showing in which tissues a specific gene is known to be expressed. The tissue expression function is only available in WebGestalt of all the tools reviewed here.

In Huang *et al.* [16] a similar data extraction and integration tool called DAVID (The Database for Annotation, Visualization and Integrated Discovery) is reported. Similarly to WebGestalt, DAVID offers functionality for analysing gene lists from different biological perspectives by providing a suite of integrated tools. The latest version of the DAVID Knowledge base consists of five integrated, web-based functional annotation features: (i) Gene Functional Classification Tool, (ii) Functional Annotation Tool, (iii) Gene ID Conversion Tool, (iv) Gene Name Viewer, and (v) NIAID Pathogen Genome Browser. This knowledge base extracts data from major well-known public bioinformatics resources by a single-linkage method which agglomerates tens of millions of diverse gene/protein identifiers and annotation terms from a variety of data sources.

Both the WebGestalt and the DAVID tools are DE methods according to our classification, as they do not perform any automated or semi-automated actual integration of the extracted data. Nevertheless, they provide a convenient organization and visualization of the extracted data, supporting the human user when manually conducting a data integration process.

Data integration methods

Von Mering *et al.* [17] proposed a comprehensive tool named STRING (Search Tool for the Retrieval of Interacting Genes/Proteins), for construction of protein-protein interaction networks. The method aims to collect, predict and unify all types of protein-protein associations. STRING integrates e.g. known protein-protein interactions, predicted interactions and results from text mining into a single protein interaction network. For organisms that have not yet been studied experimentally, STRING uses predictions and transfers known interactions from related model organisms.

Liu *et al.* [18] describe a novel method named IAB, which stands for Integration of ANN (Artificial Neural Networks) and BLAST (Basic Local Alignment Search Tool) [11]. The IAB method was developed to efficiently identify gene-specific oligonucleotides (oligos), which are short stretches of DNA used in different forms of analysis where the expression levels of genes are measured. For all such techniques the specificity of the oligos is of critical importance for the accuracy of the results. Another example where the result is highly dependent on the specificity of oligos is in the siRNA technique, which is used for silencing of genes in different biological experiments. If the oligo that is used for silencing one gene is not specific to that gene, but also binds to other gene sequences, there is a risk that the biological system is perturbed in an uncontrolled manner. Thus, one common problem with techniques that rely on the specificity of oligos is the false positives that result from cross-hybridization between highly similar sequences. This motivates the effort to improve identification of highly unique oligos, which is addressed in IAB. In the IAB method, input vectors are created for training an ANN and these are subsequently verified by BLAST.

According to the definition of DI methods in section 2, both STRING and IAB are here classified as DI methods. STRING uses heterogeneous data from several databases to derive a protein interaction network including linked associated information, for a list of input identifiers. Each data source is first used separately and the results are then integrated in the final output from the method. The IAB method uses data from two sources to improve an existing model. The data sources are used individually, as the ANN results are produced separately and then verified with results from BLAST.

Data fusion methods for heterogeneous biological data

A key issue in DF of heterogeneous biological data is that there is no gain from fusion if no increase in understanding of the underlying biological system is achieved. As discussed in [7], the results of IF should provide effective support for humans in their work to make decisions. Therefore, prior to DF it is critical to extract biologically meaningful features and detect the biologically relevant correlations between these features [19]. When inferring biological networks we work both with methods for assigning functional annotation to each of the elements in the network and with methods for analyzing causal effects between those elements. These tasks require use of data from various data sources by means of appropriate methods [2]. Here we report on selected methods for network inference [1, 2, 5], pathway reconstruction [19, 20], binding prediction [21] and functional annotation [22], all of which can be considered as examples of methods from the DF category. To certify that we have a representative selection of DF methods we have chosen methods that use different approaches such as correlation calculation, Bayesian network, clustering, multivariate regression, machine learning, combinatorial analysis, and statistical tests in the DF process.

In a study by Yamanishi *et al.* [19] a novel method was proposed for detection of correlation between three different biological data sources (functional data, geometrical data and co-expression data) with the aim of reconstructing pathways. The authors also emphasized the importance of first detecting biologically relevant correlations before fusion of data. The proposed approach consists of a generalized kernel canonical correlation analysis (KCCA) and a method for extraction of groups of genes responsible for the detected correlations.

Lee and co-workers [2] developed an IF software platform called BioCAD, consisting of three major functional modules: (i) data preprocessing, (ii) network inference, and (iii) network analysis. The platform utilizes local and global optimization for bio-network inference, text mining techniques for network validation and annotation, and infers biological networks by using various information sources.

Myers and co-workers [5] proposed a novel approach for the discovery of biological networks by merging diverse functional genomic data, such as protein-protein interactions, gene expression data, cellular localization, and results from regulation studies. Their method is implemented in a system called bioPIXIE (biological Process Inference from eXperimental Interaction Evidence) where the user can enter a set of proteins and use a probabilistic search algorithm to predict a process-specific network. The method uses a common Bayesian framework to learn from proteins or genes that are known to be functionally related. In addition, their approach for network detection also incorporates expert knowledge in the search process. The authors showed that combining data from multiple data sources clearly improves network discovery. The bioPIXIE system relies on four components: (i) Bayesian integration of heterogeneous data, (ii) an expert-driven search paradigm, (iii) a probabilistic graph search algorithm, and (iv) an interface for interpretation of the results. A limitation of the system is that it has presently only been implemented and evaluated for *Saccharomyces cerevisiae* (yeast) and not for any higher eukaryotes. Even if the applicability of this method is proposed also for higher eukaryotes, for which functional genomic data is available, the incompleteness of data is a major obstacle.

Hwang *et al.* [1] described a more general DF methodology for systems biology, named POINTILLIST. This method combines p-values from multiple data sets by using weighted versions of statistical tests, such as Fishers's exact test and Stouffer's weighted Z-method. An advantage of this method is that it is not limited to any specific organism. In addition, it can handle multiple data sets differing in statistical power, type, size, and network coverage, without requiring a curated training set. The methodology uses an optimization algorithm to minimize the number of false positives and false negatives and makes no assumptions about the number of data sets that are used in the DF process. The methodology was evaluated using simulated data sets and by trying to recapitulate a known network in yeast. The final outcome of the DF procedure is a network model where nodes represent genes or proteins, while edges represent interactions, such as transcriptional regulation.

Troyanskaya *et al.* [22] described a probabilistic framework named MAGIC (Multisource Association of Genes by Integration of Clusters) which performs fusion of heterogeneous data for gene function prediction. MAGIC incorporates protein-protein interactions, information about experimentally verified binding sites, and results from three widely used gene expression clustering methods: K-means, self-organizing maps and hierarchical clustering. The main part of MAGIC is its Bayesian network component, which combines evidence from different data sets and calculates a probability for a functional relationship between each pair of genes. MAGIC is flexible regarding the addition of new input data sets, but an essential restriction is that the method has currently only been implemented for yeast.

In Chen *et al.* [21] a computational model was proposed that fuses multiple sources of data for prediction of genes that either bind to or are regulated by the *MYC* gene in vivo. *MYC* is a regulating gene that plays a critical role in cell proliferation, growth, apoptosis, and differentiation. Many human cancers are associated with this gene, which is predicted to regulate the expression of hundreds of target genes. It is therefore important to understand the MYC binding sites and target genes to reveal the biological roles and molecular mechanisms of *MYC* action [21]. The described method uses a Bayesian network classifier and incorporates genomic sequences, experimentally determined genomic chromatin acetylation islands, and it predicts methylation status from a computational model. Based on these data, the likelihood of genomic DNA methylation is estimated. The resulting binding probability is then combined with gene expression data from a large number of microarray experiments which are conducted on various tissues. Finally, it also incorporates functional annotation data from Gene Ontology to predict target genes of MYC.

In Kasturi *et al.* [20] the authors proposed a method for fusion of gene expression data and promoter sequence data to improve the results of inferring gene regulation mechanisms at a genomic scale. A machine learning approach to DF was introduced, in which heterogeneous genomic data are combined. Studies combining sequence data with gene expression data are motivated by the idea that genes with expression profiles of similar shapes are likely to be co-regulated, and therefore might share similar sequence elements in their promoter regions [20]. This idea is also supported by results presented by Synnergren *et al.* [23]. The DF method proposed by Kasturi and co-workers uses gene expression data and binding motif data, to explore and identify relations between gene expression patterns and the presence of binding sites for transcription factors. The gene expression patterns are grouped by unsupervised clustering, using self-organizing maps. The method also takes into consideration multiple occurrences of binding sites in the upstream region of the DNA sequence.

Another study by Zhou *et al.* [24] focused on the need to fuse microarray data from multiple experiments, in contrast to fusion of diverse heterogeneous data. The increasing amount of microarray data requires an efficient way to fuse data from several experiments into one common understanding about the studied biological system. The authors proposed a novel approach to gene expression analysis, which searches for pairwise correlations between genes in microarray experiments and then compares correlation values across several experiments to derive regulatory networks. The main advantage of this method is that it is sensitive to gene pairs which are only co-expressed in a subset of the available data sets.

Similarly, the study presented in [25] focuses on fusion of multiple sets of microarray data using regression. Regression models have previously

Similarly, the study presented in [25] focuses on fusion of multiple sets of microarray data using regression. Regression models have previously been used for a variety of purposes in the analysis of microarray data, and the method proposed by Gilks and co-workers [25] is based on multivariate regression. The aim of the method is to fuse results from several microarray experiments, but with a slightly different purpose than the work presented by Zhou *et al.* [24]. The aim of the method is not to infer causes or mechanisms that underlie the data, but, equally important, to fuse data in a way that takes into account differences in the data quality. By weighting of attributes the method allows for suppression of unwanted sources of variation and delivers a fused and less biased dataset for further analysis. Even though the purpose of this method is not strictly reconstruction of biological networks, it can still be classified as a DF method according to the definitions proposed in section 2. The method clearly uses a combination of data from multiple data sets in the computation of the resulting output.

It is commonly known that the regulation of gene expression in eukaryotes is highly complex and combinatorial in its nature. This phenomenon is addressed in the DF method described by Pilpel *et al.* [26], which uses combinatorial analysis of promoter elements to uncover novel functional motifs or combinations of motifs in the promoters in the yeast genome. The method uses known and putative motifs and scans the whole genome for genes containing each motif in their promoter regions. For each motif and combination of motifs, an expression score is calculated to measure the overall similarity (across different experimental conditions) of the expression profiles for all the genes that contain that specific motif or set of motifs. Results from this method indicate that a small set of transcription factors is responsible for a large set of expression patterns under diverse conditions. A considerable limitation is that the method is only implemented for yeast, but the authors propose that the approach may be useful also for modelling the transcriptional regulatory networks of more complex eukaryotes.

Results

Methods for IF in systems biology have been reviewed with a focus on reconstruction of networks that control a particular biological process and investigation of interdependencies between genes and proteins that interact in the biological network. Furthermore, three different categories of tasks that these methods address are identified and illustrated in Fig. 1. Each one of the reviewed IF methods has been classified into one of the three categories DE, DI, and DF, as presented in Tab. 1.

Table 1: Classification of methods.

Method name[a]	Reference[b]	Category[c]	Restriction to protein or organism[d]	Type of data[e]
WebGestalt	[3]	DE	No	Various
DAVID	[16]	DE	No	Various
IAB	[17]	DI	No	Sequences
STRING	[12]	DI	No	Genes Proteins
KCCA	[19]	DF	No	Function Geometry Co-expression
BioCAD	[2]	DF	No	Various
bioPIXIE	[5]	DF	Yes	Protein interaction Gene expression Localization Regulation
POINTILLIST	[1]	DF	No	Various
MAGIC	[22]		Yes	Protein interaction Binding sites Gene clusters
N/A	[21]	DF	Yes	Sequences Chromatin acetylation islands GO annotation
N/A	[20]	DF	No	Motifs Gene expression
N/A	[24]	DF	No	Gene expression
N/A	[25]	DF	No	Gene expression
N/A	[26]	DF	No	Motifs Gene expression

[a] Name of the method (if it has a name, otherwise N/A).
[b] Reference which describes the method.
[c] One of the three proposed categories Data Extraction (DE), Data Integration (DI) and Data Fusion (DF).
[d] Indicates if the method is restricted to specific proteins or organisms.
[e] Examples of the type of data the method incorporates.

Conclusion and discussion

A conclusion from the present work is that IF in systems biology is commonly conducted in two main steps: First extraction of relevant data for investigation of the biological process and, secondly, integration or fusion of the extracted data into a coherent understanding of the underlying

A conclusion from the present work is that IF in systems biology is commonly conducted in two main steps: First extraction of relevant data for investigation of the biological process and, secondly, integration or fusion of the extracted data into a coherent understanding of the underlying biological process (Fig. 1). Different methods for assisting the researcher in these steps have been reviewed with a focus on methods using multiple heterogeneous data sets when trying to reconstruct interaction networks representing an underlying biological process, as well as methods for exploring the roles of genes and proteins that interact in the process. These methods were further classified into three identified categories of tasks, as illustrated in Fig. 1. Based on this work it can also be concluded that the coverage of appropriate IF methods in systems biology varies across the identified categories of tasks. Several extensive tool-suites are freely available that fulfill the need for DE. These are well adapted for the biological needs and implemented with easy-to-use interfaces suitable for different types of users. However, when it comes to integration of heterogeneous data, very few methods for automation of this process can be found, and the user is often left to carry out the integration tasks more or less manually. This is a time-demanding and to some extent subjective work that requires extensive biological knowledge from the user, if a biologically meaningful integration of the data is to be achieved.

Regarding DF this work reviews several methods that focus on appropriate fusion of various types of biological data. A challenge is the difficulty to find DF strategies that are general enough to apply to a range of different fusion problems. Nevertheless, it should be emphasized that the bottleneck for this category is still to make a biologically relevant fusion of data that reflects the underlying biological system under study. The knowledge about relations between different types of data is limited, which is an obstacle to any attempt at developing efficient DF methods. One should always have in mind that the fusion of data should result in an improved understanding of the underlying biological system, because otherwise there is no motivation for the DF [4]. The critical question to ask when trying to gain more knowledge about a biological process is therefore, "What data should be fused and how?".

Here we presented a classification of IF-related bioinformatic methods into three categories DE, DI and DF based on the functionality of the methods. In this study a representative subset of methods for extraction, integration and fusion of heterogeneous biological data were reviewed. This subset of methods reflects the diversity of IF methods that have been developed in the systems biology community. Taken together, IF is a rapidly developing research field that can make substantial contributions to automated or semi-automated integration and fusion of large-scale heterogeneous biological data sets. Therefore, to facilitate the utilization of IF methods across research domains it is essential to have access to a uniform and standardized classification scheme, appropriate for methods from a wide range of research areas. A model-based approach might potentially be useful for the classification. One such model is the widely known JDL Data Fusion model [27], which has mainly been used in the military domain. In Synnergren *et al.* [28] it was shown that the JDL model is general enough to also apply to IF-methods for bioinformatics. The model can be used as a communication tool across communities and provides a common vocabulary. The JDL model consists of five levels of DF processes. Whether these levels will provide a coarser or more detailed classification of methods than what is presented here remains to be investigated. An extended classification of the methods presented in Tab. 1 (and possibly additional relevant IF methods) based on the five different levels in the JDL model, is therefore suggested as future work.

Acknowledgements

This work was supported by the Information Fusion Research Program (www.infofusion.se) at the University of Skövde, Sweden, in partnership with the Swedish Knowledge Foundation under grant 2003/0104, and participating partner companies.

References

1. Hwang, D., Rust, A. G. Ramsey, S., Smith, J. J., Leslie, D. M., Weston, A. D., de Atauri, P. Aitchison, J. D., Hood, L., Siegel, A. F. and Bolouri, H. (2005). A data integration methodology for systems biology. Proc. Natl. Acad. Sci. USA **102**, 17296-17301.
2. Lee, D., Kim, S. and Kim, Y. (2007). BioCAD: an information fusion platform for bio-network inference and analysis. BMC Bioinformatics **8 Suppl. 9**, S2.
3. Zhang, B., Kirov, S. and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. **33**, W741-W748.
4. Hall, D. L. and Llinas, J. (1997). An Introduction to Multisensor Fusion. Proceedings of the IEEE **85**, 6-23.
5. Myers, C. L., Robson, D., Wible, A., Hibbs, M. A., Chiriac, C., Theesfeld, C. L., Dolinski, K. and Troyanskaya, O., G. (2005). Discovery of biological networks from diverse functional genomic data. Genome Biol. **6**, R114.
6. Merelli, I., Morra, G., D'Agostino, D., Clematis, A. and Milanesi, L. (2005). High performance workflow implementation for protein surface characterization using grid technology. BMC Bioinformatics **6 Suppl. 4**, S19.
7. Boström, H., Andler, S. F., Brohede, M., Johansson, R., Karlsson, A., van Laere, J., Niklasson, L., Nilsson, M., Person, A. and Ziemke, T. (2007). On the definition of information fusion as a field of research. Technical report at the University of Skövde, HS-IKI-TR-07-006.
8. MSN Encarta. Data fusion definition. URL: http://encarta.msn.com/dictionary_/data%2520fusion.html (Accessed February 2008).
9. Lenzerini, M. (2002). Data Integration: A Theoretical Perspective. PODS, 233-246.
10. Abidi, M. A. and Gonzalez, R. C. (1992). Data Fusion in Robotics and Machine Intelligence. Academic Press, San Diego.
11. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**, 403-410.
12. Montaner, D., Tárraga, J., Huerta-Cepas, J., Burguet, J., Vaquerizas, J. M., Conde, L., Minguez, P., Vera, J., Mukherjee, S., Valls, J., Pujana, M. A. G., Alloza, E., Herrero, J., Al-Shahrour, F. and Dopazo, J. (2006). Next station in microarray data analysis: GEPAS. Nucleic Acids Res. **34**, W486-W491.
13. Al-Shahrour, F. Minguez, P., Tárraga, J., Montaner, D., Alloza, E., Vaquerizas, J. M., Conde, L., Blaschke, C., Vera, J. and Dopazo, J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res. **34**, W472

13. Al-Shahrour, F. Minguez, P., Tárraga, J., Montaner, D., Alloza, E., Vaquerizas, J. M., Conde, L., Blaschke, C., Vera, J. and Dopazo, J. (2006) . BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res. **34**, W472-W476.

14. Ogmen, U., Keskin, O., Aytuna, A. S., Nussinov, R. and Gursoy, A. (2005). PRISM: protein interactions by structural matching. Nucleic Acids Res. **33**, W331-W336.

15. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**, 25-29.

16. Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C. and Lempicki, R. A. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. **35**, W169-W175.

17. von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B. and Bork, P. (2007). STRING 7-recent developments in the integration and prediction of protein interactions. Nucleic Acids Res. **35**, D358-D362.

18. Liu, C.-C., Lin, C.-C., Li, K.-C., Chen, W.-S., Chen, J.-C., Yang, M.-T., Yang, P.-C., Chang, P.-C. and Chen, J. J. W. (2007). Genome-wide identification of specific oligonucleotides using artificial neural network and computational genomic analysis. BMC Bioinformatics **8**, 164.

19. Yamanishi, Y., Vert, J.-P., Nakaya, A. and Kanehisa, M. (2003). Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics **19 Suppl. 1**, i323-i330.

20. Kasturi, J. and Acharya, R. (2005). Clustering of diverse genomic data using information fusion. Bioinformatics **21**, 423-429.

21. Chen, Y., Blackwell, T. W., Chen, J., Gao, J., Lee, A. W. and States, D. J. (2007). Integration of genome and chromatin structure with gene expression profiles to predict c-MYC recognition site binding and function. PLoS Comput. Biol. **3**, e63.

22. Troyanskaya, O.G., Dolinski, K., Owen, A. B., Altman, R. B. and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc. Natl. Acad. Sci. USA **100**, 8348-8353.

23. Synnergren, J. Adak, S., Englund, M. C., Giesler, T. L., Noaksson, K., Lindahl, A., Nilsson, P., Nelson, D., Abbot, S., Olsson, B. and Sartipy, P. (2007). Cardiomyogenic gene expression profiling of differentiating human embryonic stem cells. J. Biotechnol. **134**, 162-170.

24. Zhou, X. J., Kao, M.-C. J., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O. M., Finch, C. E., Morgan, T. E. and Wong, W. H. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. Nat. Biotechnol. **23**, 238-243.

25. Gilks, W. R., Tom, B. D. M. and Brazma, A. (2005). Fusing microarray experiments with multivariate regression. Bioinformatics **21 Suppl. 2**, ii137-ii143.

26. Pilpel, Y., Sudarsanam, P. and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. Nat. Genet. **29**, 153-159.

27. Steinberg, A. N., Bowman, C. L. and White, F. E. (1999). Revisions to the JDL data fusion model. SPIE Conference on Sensor Fusion: Architectures, Algorithms, and Applications III. Orlando, Florida, USA, pp. 430-441.

28. Synnergren, J., Gamalielsson, J. and Olsson, B. (2007). Mapping of the JDL data fusion model to bioinformatics. Systems, Man and Cybernetics, 2007 IEEE International Conference, Montreal, QC, Canada, pp. 1506-1511.