NETWORK-BASED REPRESENTATION OF BIOLOGICAL DATA FOR ENABLING CONTEXT-BASED MINING

Catherine Bounsaythip[1], Erno Lindfors[1], Peddinti V. Gopalacharyulu[1], Jaakko Hollmén[2], Matej Orešič[1]

[1]VTT Biotechnology,
P.O. Box 1500, Espoo, FI-02044 VTT, Finland, *name.surname*@vtt.fi, ext-Gopal.Peddinti@vtt.fi,
[2]Helsinki University of Technology,
Laboratory of Computer and Information Science, P.O. Box 5400, Espoo, FIN-02015 HUT, Finland, Jaakko.Hollmen@hut.fi

ABSTRACT

Biological phenomena are usually described by relational model of interactions and dependencies between different entities. Therefore, a network-based knowledge representation of biological knowledge seems to be an obvious choice. In this paper, we propose such a representation when integrating data from heterogeneous life science data sources, including information extracted from biomedical literature. We show that such a representation enables explanatory analysis in a context dependent manner. The context is enabled by a judicious assignment of weights on the quality dimensions. Analysis of clusters of nodes and links in the context of underlying biological questions may provide emergence of new concepts and understanding. Results are obtained with our *megNet* software, an integrative platform based on a multi-tier architecture using a native XML database.

1. INTRODUCTION

The primary goal of knowledge representation is to enable computer to assist humans in analyzing complex forms of data to discover useful information. This has resulted in a wide range of techniques and tools. How to represent knowledge depends largely on the way reasoning can be done with that knowledge. For example, early works have been mainly focused on logic-based representation. Recently, techniques combining machine learning, pattern recognition, statistics, and artificial intelligence have been employed. Although these are well-developed disciplines, their applications in life science have been limited [1][2][3].

Biology is a data rich discipline. The problem is that this source of knowledge is stored in a large number of different data sources which need to be mined in parallel. Integrating all this information and its efficient mining is a challenge with huge application potential [4][5]. Moreover, each database may have its own interface that users may not have time to adequately learn to use them efficiently. A tool which can integrate the mining as well as visualization of heterogeneous life science data would therefore open new possibilities for the exploration of biological knowledge and possibly lead to novel discoveries.

As biological systems are characterized by the complexity of interactions of their internal parts and also with the external environment, integrating such interacting information may result in a large connected graph with nodes and edges of heterogeneous types. This makes such information hard to visualize, and sophisticated methods have been developed for analyzing such complex networks [6][7][8][9]. The most important aspect in visualizing high-dimensional data in a lower dimensional space is how to preserve the proximity relationships. In practice, it is very difficult if not impossible to project hundreds of dimensional data to a smaller dimensional space (2 or 3 dimensions) in such a way that all similarity relationships are preserved. Therefore, in order to enable effective reasoning, the challenge is to find the best compromises by choosing which kinds of relationships to visualize and with what type of metrics to use in order to ensure the trustworthiness of the visualized data [10].

Another way to enable effective reasoning is to limit the scope of deliberations to a small context associated with the domains under consideration. This may be approached by assigning weights to the "quality dimensions" [11] under consideration (gene-centric, tissue-centric, compound-centric, disease-centric etc.)

The above criteria have been our motivations to develop an integrated visualization tool, *megNet*, that uses topological analysis of complex networks to visualize query results in a single interface. It also enables context-based information display from our integrated database system (see [12]).

This paper discusses the representation and visualization aspects of our integration platform. It is organized as follows: Section 2 discusses about the network representation and clustering methods, including the notion of distance and context. Section 3 gives examples of visualizing a protein-protein interaction network.

2. BIOLOGICAL NETWORKS

With the growing trend towards systems biology, integrated biological networks contain many different types

of entities and attributes arising from a growing number of disparate data sources, including literature databases. These databases have been created by different scientific communities, for different purposes, and covered different aspects. All that led to a high level of structural and semantic heterogeneity. The structural and semantic integration aspects of these databases have been reported in our previous papers [12][13]. Here we will focus on the retrieval and visualization of these heterogeneous data. We are mainly interested in the data from the following databases:

- Protein-protein interaction databases: *BIND* [14], *DIP* [15], and *MINT* [16].
- Biochemical pathways database: *KEGG* [17].
- *TransFac* is a database on DNA binding elements and their transcription factors [18].
- *TransPath*, an extension of *TransFac*, contains signal transduction pathways that regulate the activity of transcriptional factors in different species [19].
- *GeneOntology* (GO) is a database of three structured controlled vocabularies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner [20].

The first step after retrieving all the massive information from databases is to build the network. The objects in network are then clustered based on some similarity measure for the display. The definition of the similarity measure is thus a crucial step.

2.1. Network representation

The graph representation contains nodes and edges [21][22]. The nodes include various kinds of molecules, e.g., proteins, compounds, genes, mRNAs etc. For example, in the case of protein-protein interaction network, we would relate the neighboring proteins by searching all the possible pathways among them, including their regulating genes. The generated nodes and edges show the proteins and their interactions, respectively.

Our biological network is presented as a directed weighted graph where biological entities are nodes that are connected to each other through edges which are interactions between the entities. The shape of the nodes will be coded differently depending on the type of an entity. The edges can be directed or undirected depending on the nature of the interactions (Figure 1).

A metabolic network consists of *reactions*. In one reaction there are *substrates*, *products* and at least one *enzyme* that catalyzes the reaction. The substrates, products and enzymes are presented as nodes. The substrates and products are presented as circles and the enzymes are presented as squares. Since some reactions are reversible and other reactions are irreversible, directed edges are used to distinguish the direction of a reaction. But in a protein-protein interaction network, interactions between the proteins are represented with undirected edges, because the interaction is mutual.

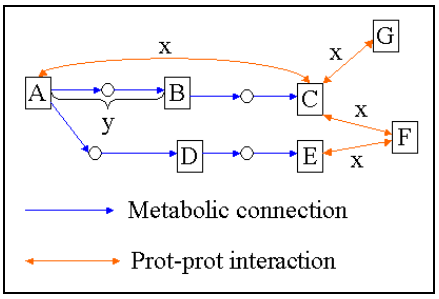

Figure 1: Example of our integrated network representation used. The distance between the entities A and B, is the same as for B to A. If there is not any path between two nodes, we assume that the distance between them is infinity.

The shortest path length between each entity is obtained by using Tom Sawyer Java analysis toolkit (Tom Sawyer, Inc.). The distances between each entity in both directions are calculated, based on the cost of connection types. In Figure 1, the cost of a metabolic interval is denoted by *y*, and *x* is the cost of a protein-protein interaction. By changing these cost parameters we can investigate how protein-protein interactions affect the structure of metabolic pathways.

2.2. Clustering of biological networks

The molecular entities of the cell form a very complicated and dynamic interacting system. Yet, it has been demonstrated that this complex interactions shared some common network properties, e.g. the presence of single modularity networks [24][25][26]. However, the presence of the modularity in highly integrated biological networks is not self-evident as it lacks quantitative support [24]. There is thus a need for tools to identify the modularity of a biological network and to identify the modules and their relationships. Clustering is a mathematical method which allows the identification of key connectivity patterns of a network. The most common methods used when investigating the structure of complex networks are hierarchical clustering tree, Kohonen's Self-Organizing Maps (SOM) [28], and Sammon's mapping [29][30].

All clustering algorithms share the basic steps:

1. *Compute distance matrix;*
2. *Find closest pair of clusters;*
3. *Update distance matrix.*

First, the distance matrix must be computed. The distance matrix define distances from one entity to the other entities. The distance matrix from the graph represented in Figure 1 is:

$$D=\begin{bmatrix} 0 & y & \min(x,2y) & y & 2y & \min(x,2y)+x \\ 3y+x & 0 & y & \inf & y+2x & y+x \\ x & x+y & 0 & x+y & \min(x+2y,2x) & x \\ y+3x & 2y+2x & y+2x & 0 & y & y+x \\ 3x & 3x+y & 2x & 3x+y & 0 & x \\ 2x & 2x+y & x & 2x+y & x & 0 \end{bmatrix}$$

If the purpose of the distance calculations is to investigate the structure of metabolic pathways, the distance matrix would not take into account metabolites and other

proteins that do not belong to the metabolic pathway (e.g. entities F and G in Figure 1).

After the distance matrix has been obtained, we can apply clustering algorithm which will merge objects in the same cluster based on the self-similarity. The self-similarity of a group of elements is defined as the average pairwise similarity between the elements. One may also choose other criteria such that the pair of clusters maximizes the minimum similarity or minimize the maximum similarity.

Since the purpose of the distance matrix is to describe the proximity of the entities, the more similar distance vectors are, the closer are corresponding biological entities. In our current implementation, we use the Sammon's mapping algorithm to investigate the similarities of the distance vectors.

2.2.1. Similarity measure

For integrated network where entities are of complex nature, evaluating similarity is not a trivial task. While distances within the molecular networks can be intuitively set to the length of the shortest path between the molecules, distance measure is less obvious for relationships such as in ontologies. It was shown that GeneOntology can be represented as a graph, and the distance measures based on the shortest path to a common ancestor were already studied [31]. In the case of gene expression network which consists only of genes, the similarity measure is based on the gene expression level.

The challenge is to combine topology metrics and the quantitative information from the data. For instance, one can combine the gene expression level and the topology of the network in the same distance function such as in [32]: $d = f(\delta_{exp} + \delta_{net})$.

Given a set of data points x_i, let us note by $d(x_i, x_j)$ being the distance between two data points.

If we consider the gene expression level G_{ik} as a log-ratio gene expression of gene $g_{i,}$, the distance function could be based on the Pearson correlation coefficient:

$$\rho_{exp}(g_i, g_j) = \frac{1}{N}\sum_k \left(\frac{G_{ik} - \mu_i}{\sigma_i}\right)\left(\frac{G_{jk} - \mu_j}{\sigma_j}\right)$$

with μ_i and σ_i are mean and standard deviation of the transformed time series data of g_i.

The correlation coefficient is then converted to a distance function as a degree of dissimilarity with: $\delta_{exp}(g_i, g_j) = 1 - \rho(g_i, g_j)$. We obtain the combined distance function:

$$d(x_i, x_j) = 1 - 0.5 \times (\delta_{exp}(g_i, g_j) + \delta_{net}(v_i, v_j)$$

The network distance function could be based on the shortest path and the weighting function based on the degree of vertices.

It is supposed that this combined function may lead to increased stability of clustering solution when the gene expression levels support the relations in the networks and vice versa [32].

In our current implementation, gene expression databases are not yet fully operational for integrated mining.

2.2.2. Data projection and non-linear mapping

The main purpose of data projection is to transform a high dimensional data to a lower dimensional space in order to be able to visualize them. The Kohonen's self-organizing map (SOM) [28] is one popular method. But the delicate part of SOM is that the user needs to set control parameters carefully that may require sometimes *a priori* knowledge about the data. We have chosen the Sammon's mapping [29] as is easier to implement.

Like the SOM algorithm, the basic idea of the Sammon's mapping algorithm is to arrange all the data points on a 2-dimensional plane in such a way, that the distances between the data points in this output plane resemble the distances in vector space as defined by some metric as faithfully as possible. Unlike SOM algorithm, the Sammon's mapping algorithm tries to preserve internal distances in the input data that the human eye can easily detect. The structure of the input data is thus preserved through the mapping.

More formally, let d_{ij} be an element of a distance matrix D in input space, let o_i be the image of the data item x_j in the 2-dimensional output space. With O we denote the distance matrix containing the pairwise distances between images as measured by the Euclidean vector norm $\|o_i - o_j\|$. The goal is to place the o_i in such a way that the distance matrix O resembles as closely as possible matrix D, i.e. to optimize an error function E by following an iterative gradient-descent process:

$$E = \frac{1}{\sum_i \sum_{j>i} d_{ij}} \sum_i \sum_{j>i} \frac{(d_{ij} - \|o_i - o_j\|^2}{d_{ij}}$$

The resulting visualization depicts clusters in input space as groups of data points mapped close to each other in the output plane. Thus, the inherent structure of the original network can be derived from the structure detected in the 2-dimensional visualization.

2.3. Context

When a representation includes several domains, one must take into account the context in which what domains appear more or less important (or *salient*) [9].

Including context can be achieved by assigning *weights* to each domain. The relative weight of a domain will depend on the context.

2.3.1. Weights as context dependent variables

In the previous section, the distance function could be weighted as follows:

$$D_{ij} = \sum_{k=1}^{n} w_k d_{ijk}$$

The weights w_k can be seen as *context-dependent* variables that represent the relative degree of salience for each dimension. This aspect has been used in the subspace clustering algorithms which assume that cluster may exist in different subspaces of different sizes. For example, in the COSA algorithm [33], the weights are assigned to each dimension for each instance, not each

cluster. Higher weights are assigned to those dimensions that have a smaller dispersion within the *k-nearest* group. The neighborhoods for each instance become iteratively enriched with instances belonging to its own cluster. The dimension weights are refined as the dimensions relevant to a cluster receive larger weights. This process enables some dimensions to emerge by different the clustering criteria. However, in the COSA algorithm, the number of dimensions to be included in a cluster cannot be set directly by the user, it is done through a parameter λ, which controls the incentive for clustering on more dimensions.

This COSA distance was shown to be more powerful than traditional Euclidean distance.

Therefore, the choice of the similarity measure can affect greatly the quality of the visualization in the projection space. When we change dimension in the visualization, the degree of similarity between two data points changes with the salience of the dimensions of the objects. This aspect was investigated in [9].

It must be noticed also that the knowledge and interest of the user may influence the "salience weights" as it is assumed that people can have different "perspectives". Therefore it is important that the user has also the possibility to influence this parameter in the visualization tool.

2.3.2. The effect of context in knowledge discovery

With the explosion of information resources on the Web, ontologies have been extensively developed to facilitate the understanding, sharing, re-use and integration of knowledge through the construction of an explicit domain model. In life science, the efforts in building ontologies across domains still have many challenges to go through [34][35]. Gene Ontology (GO) is the only ontology that has been extensively used in bioinformatics [36][37]. However, GO seems to be more a taxonomy rather than a well-formed ontological structure that would enable traditional rule-based reasoning [38]. Another drawback of GO and other Ontologies in general, is their static structure and thus, when used as a structure for reasoning, they can only produce *monotonic* inference. Such a mode of reasoning may hinder or possibly even prevent the discovery and exploration of new possibilities [39].

While in a context-based reasoning, the conceptualization associated to the "cluster" that has emerged from the context, is *non-static*. For example, when we interpret clusters obtained from gene expression data, we must take into account the context of underlying biological models e.g., from which tissue and what was environmental history which has led to that state.

3. EXAMPLES

In this section we would like to give an example of network clustering of data retrieved from metabolic pathways and protein-protein interaction databases. As an example, we create a network based on the KEGG metabolic pathway from the query: "Glycolysis / Gluconeogenesis, Pentose phosphate and Citrate cycle pathways", for *S. cerevisiae* (Figure 2). The enzymes are then enriched with protein-protein interaction (MINT, DIP).The query results are shown in Figure 3. We can see from the Sammon's mapping that there are two main clusters in these pathways, a strongly connected cluster and sparsely connected cluster (Figure 3). Sparsely connected proteins are highlighted with gray marks, which appear to be mostly located at the border of the graph. Based on the concept of hierarchical modularity, we may conclude that the proteins of the strongly connected cluster are in higher hierarchy level than those of the sparsely connected cluster.

Another example of search is performed for protein-protein interaction with the set of proteins {P41940, O15305, P29952} which are involved in the glycosylation and mannosylation pathways in S. *cerevisiae*, referenced in GeneOntology Biological process "GDP-mannose biosynthesis" with GO:0009298. Results are shown in Figure 5. Clustering examples with different contexts (different weight assignments) are given in Figure 6 and Figure 7. In Figure 6, all the edges have equal weights. We can see that he neighborhood of GO:0009298 consist of proteins C05345 and C00275, which denote that in this context, they have stronger connection to GO:0009298. In Figure 7, the neighbors of GO:0009298 have larger weights, this has resulted in the clustering of proteins of the query set {P41940, O15305, P29952}.

We can "experiment" with the weight assignment for different context and notice that relative proximity of nodes changes. This might suggest new hypotheses that these entities might be involved in the same process or pathways reflected by the context.

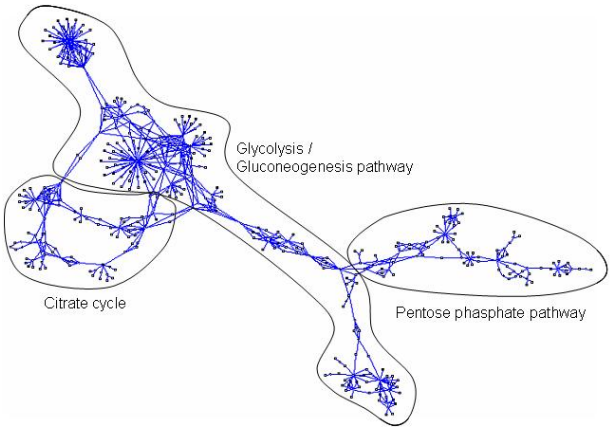

Figure 2: KEGG metabolic pathways for "Glycolysis / Gluconeogenesis , Pentose phosphate and Citrate cycle pathways.

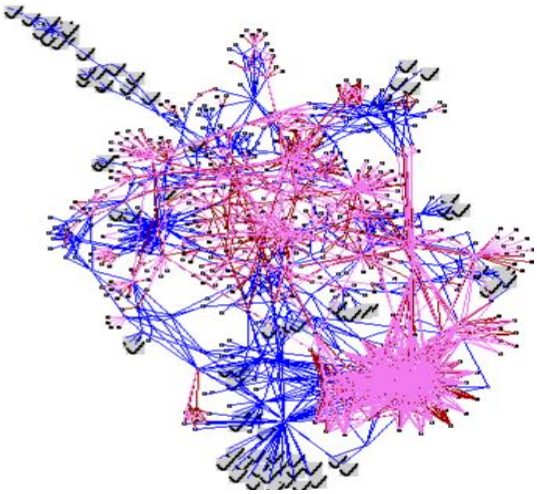

Figure 3: Metabolic pathway (KEGG) enriched with protein-protein interactions from MINT and DIP databases for “Glycolysis / Gluconeogenesis, Pentose phosphate and Citrate cycle pathways,. The proteins loosely connected are highlighted with gray marks.

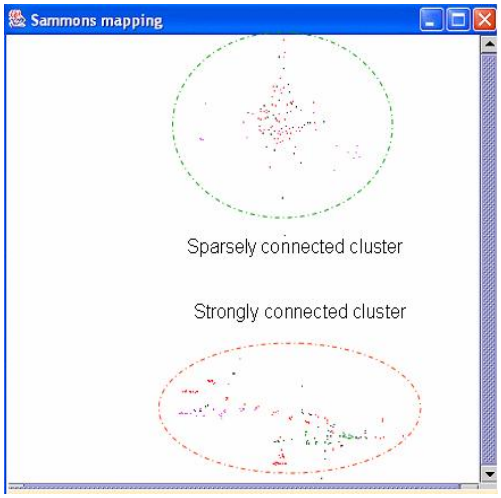

Figure 4: Clusters from Sammon’s mapping of the previous graph. Two main clusters emerged, one strongly connected and one loosely connected.

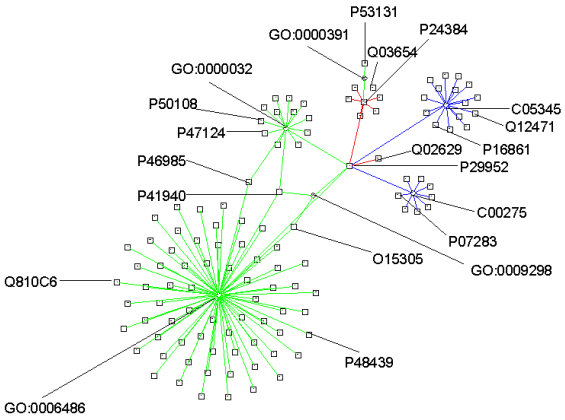

Figure 5: Search result of pathway query for mannose synthesis GO:0009298.

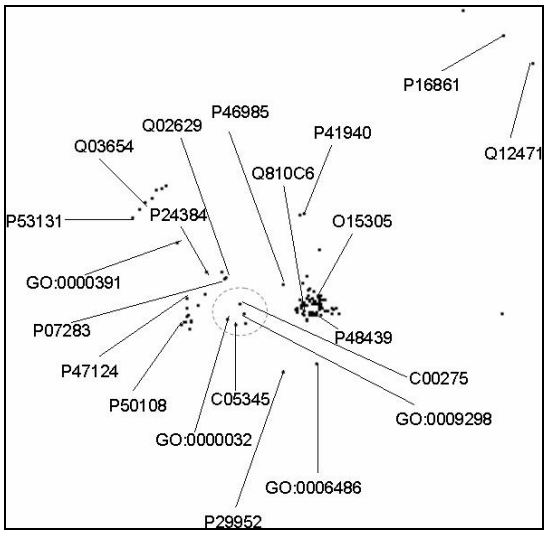

Figure 6: Sammon’s mapping of the previous network for ”Context 1: Every edge has equal weight”.

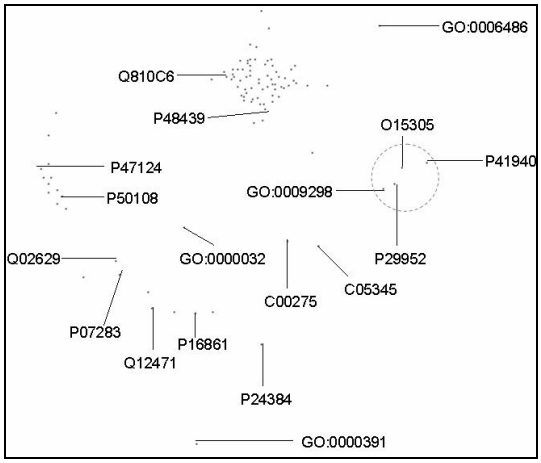

Figure 7: The Sammon’s mapping for “Context 2: The neighborhood edges of GO:0009298 have higher weights than the other edges”.

4. CONCLUSION

In this paper we have discussed about the heterogeneity of biological data and resources and existing methodologies to analyze those data. We introduced our approach to represent integrated biological data for enabling visual exploratory analysis. At the current phase, we have implemented the Sammon’s mapping clustering with a distance function that incorporates the notion of context, which can be controlled by the user. Our experiments have shown that the Sammon’s mapping algorithm is not very suitable for a large number of input vectors. Therefore, in our biological networks consisting of a large number of nodes, clustering time is rather long. Second, one cannot always rely totally on the output by the Sammon’s mapping clustering due to the trustworthiness of distance function. Therefore, it is up to the user to look for insight and experiment with the dimension salience to see if it makes any sense and always reconnect to the original hypothesis and background knowledge.

5. REFERENCES

[1] F. Capra, *The Web of Life*, Harper Collins, London, 1997.

[2] D. B. Kell, S. G. Oliver, " Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era", *Bioessays*; 26(1), pp. 99-105, 2004.

[3] F. Katagiri. "Attacking Complex Problems with the Power of Systems Biology", *Plant Physiology*, Vol. 132, pp. 417–419, 2003.

[4] D. B. Searls, "Data integration: challenges for drug discovery". *Nature Reviews Drug Disc*., 4, pp. 45-48, 2005.

[5] R. B. Stoughton, S. H. Friend, "How molecular profiling could revolutionize drug discovery", *Nature Rev. Drug Disc.*, Vol. 4, pp. 345-350, 2005.

[6] H. Jeong, B. Tombo, R. Albert, Z.N. Oltvai, A.-L. Barabási, "The Large-Scale Organization of Metabolic Networks", *Nature*, vol. 407, p. 651, 2000.

[7] M. E. J. Newman "The structure and function of complex networks", *SIAM Review*, 45(2), pp. 167- 256, 2003.

[8] A.-L. Barabási and Z. N. Oltvai, "Network Biology: Understanding the Cells' Functional Organization", *Nature Reviews Genetics*, vol. 5, pp. 101-114, Feb. 2004

[9] J. A. Papin, B. O. Palsson, "Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk". *J. Theor. Biol.*, 227, pp. 283-297, 2004.

[10] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen and E. Castrén, "Trustworthiness and metrics in visualizing similarity of gene expression", *BMC Bioinformatics*, pp. 4-48, 2003.

[11] P. Gärdenfors, *Conceptual spaces: The geometry of thought*, MIT Press, Cambridge, MA, 2000.

[12] P.V. Gopalacharyulu, E. Lindfors, C. Bounsaythip, T. Kivioja, L. Yetukuri, J. Hollmén, and M. Orešič, "Data integration and visualization system for enabling conceptual biology", *Proc. of International conference on Intelligent Systems for Molecular Biology* (ISMB 2005), Detroit, MI, USA, June 25-29, 2005.

[13] P. V. Gopalacharyulu, E. Lindfors, C. Bounsaythip, W. Wefelmeyer & M. Orešič, "Ontology based data integration and context-based mining for life sciences", *Proc. W3C Workshop on Semantic Web for Life Sciences*, Cambridge, MA, USA, 2004.

[14] G. D. Bader , D. Betel, C. W. V.Hogue, "BIND: the Biomolecular Interaction Network Database", *Nucl. Acids Res.*, 31, pp. 248-250, 2003.

[15] The DIP database, http://dip.doe-mbi.ucla.edu/

[16] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, G. Cesareni, "MINT: a Molecular INTeraction database", *FEBS Lett.*, 513, pp.135-140, 2002.

[17] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, "The KEGG resource for deciphering the genome", *Nucl. Acids Res.*, 32, pp. 277-280, 2004.

[18] V. Matys, E. Fricke, R. Geffers, E. Gossling, *et al.* "TRANSFAC: transcriptional regulation, from patterns to profiles", *Nucl. Acids Res.*, vol. 31, pp. 374-378, 2003.

[19] M. Krull, N. Voss, C. Choi, S. Pistor, A. Potapov, E. Wingender, "TRANSPATH: an integrated database on signal transduction and a tool for array analysis", *Nucl. Acids Res.*, 31, pp. 97-100, 2003.

[20] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, "Gene ontology: tool for the unification of biology", *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[21] B. Bollobás, *Modern Graph Theory*, Graduate Texts in Mathematics, vol. 184, Springer, New York, 1998.

[22] R. Diestel, *Graph Theory*, Graduate Texts in Mathematics, vol. 173, Springer, New York, 1997.

[23] S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of Networks From Biological Nets to the Internet and WWW*, Oxford University Press, Oxford, UK, 2003.

[24] A.-L. Barabási, Z. N. Oltvai, "Network Biology: Understanding the Cells' Functional Organization", *Nature Reviews Genetics*, vol. 5, pp. 101-113, 2004.

[25] J. D. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L. V. Zhang, D. Dupuy, A.J. Walhout, M.E. Cusick, F.P. Roth, M. Vidal," Evidence for dynamically organized modularity in the yeast protein-protein interaction network", *Nature*, vol. 430, pp. 88-93, 2004.

[26] R. Guimera, L. A. Nunes Amaral, "Functional cartography of complex metabolic networks", *Nature*, vol. 433, pp. 895-900, 2005.

[27] E. Ravasz, A.-L. Barabási, "Hierarchical organization in complex networks", *Physical Review*, vol. 67, pp. 026112, pp. 1-7, 2003.

[28] T. Kohonen, *Self-Organizing Maps*, Springer Verlag, 2001.

[29] J. W. Sammon Jr., "A nonlinear mapping for data structure analysis". *IEEE Trans. Comp.*, C-18, 401-409, 1969.

[30] F. Azuaje, H. Wang, A. Chesneau, "Non-linear mapping for explanatory data anlysis in functional genomics", *BMC Bioinformatics*, pp. 6-13, 2005.

[31] S. G. Lee, J. U. Hur, Y. S. Kim, "A graph-theoretic modeling on GO space for biological interpretation of gene clusters". *Bioinformatics*, vol. 20, pp. 381-388, 2004.

[32] D. Hanisch, A. Zien, R. Zimmer, T. Lengauer, "Co-clustering of biological networks and gene expression data", *Bioinformatics*, Vol. 18, pp. 145-154, 2002.

[33] J. F. Friedman, J. J. Meulman, "Clustering objects on subsets of variables". *Journal of the Royal Statistical Society*, Series B, 4, pp. 815-849, 2004.

[34] R. Stevens, C. Wroe, P. Lord, C. Goble, "Ontologies in bioinformatics". *Handbook on Ontologies in Information Systems*, pp. 635-657, Springer, 2003.

[35] J. L. Bard, S. Y. Rhee, "Ontologies in biology: design, applications and future challenges", *Nature Review Genetics*, vol. 5(3), pp. 213-22, 2004.

[36] M. A. Harris et al. "The Gene Ontology (GO) database and informatics resource", *Nucleic Acids Res.* vol. 32 Database issue, pp. 258-261, 2004.

[37] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, R. Ap-weiler. "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology". *Nucl. Ac-ids Res.*, vol. 32, pp. 262-266, 2004.

[38] B. Smith, J. Williams, S. Schulze-Kremer, "The Ontology of the Gene Ontology", *in Proc. of the Annual Symposium of the American Medical Informatics Association*, Washington DC, Nov. 2003.

[39] C. Catton, D. Shotton, "The use of Named Graphs to enable ontology evolution", *W3C Workshop on the Semantic Web for Life Sciences*, Cambridge, MA, USA, 2004.