Analysis of integrated biomolecular networks using a generic network analysis suite

Matthias Oesterheld[1], Hans Werner Mewes[1,2], Volker Stümpflen[1]

[1]Institute for Bioinformatics, German National Center for Health and Environment, Helmhotz Association, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany
[2]Technische Universität München, Wissenschaftszentrum Weihenstephan, Lehrstuhl für Genomorientierte Bioinformatik, Am Forum 1, 85435 Freising-Weihenstephan, Germany

Abstract

The informative value of biomolecular networks has shifted from being solely information resources for possible cellular partners (whether these embody proteins, (ribo)nucleic acids or small molecules) towards becoming models for the functional connectivity within a cell. These models are increasingly exploited to make quantitative predictions about the cell's functional organization as well as about the functionality of individual elements in the network.

A large number of concepts and methods have been proposed in order to interpret experimental data mapped to cellular networks these systems and to make use of the rich source of information they represent.

We will present a system for the Comprehensive Analysis of Biomolecular Networks (CABiNet), capable of integrating available network analysis methods. Integration is done by classifying each method into one of four separate categories using standardized interfaces that encapsulate the functionality of the method in a distinct component with standardized in- and output. These components can be accessed individually or in an integrated form using a processing pipeline for semi-automatic analyses.

Additionally, the system can be used to query both biomolecular networks as well as the derived results of network analysis methods, such as clustering algorithms, in order to provide a service for researchers who are focused towards the functional context of any particular cellular entity.

CABiNet is designed in an easy-to-use and easy-to-extend software framework that allows a straightforward integration of novel components. We will demonstrate the capabilities of the system by introducing several use cases.

The CABiNet suite can be accessed at http://mips.gsf.de/genre/proj/CABiNet. Source code including additional components that can be accessed using the API is available upon request.

1 Introduction

Systems biology has shifted the focus of research from studying individual cellular entities towards comprehensive studies of their intricate connections on a system level [1]. Biological systems are functionally organized into different related networks, representing the relationship between the cellular components. Various domains of biomolecular networks exist, defined by the particular type of interaction, such as protein interaction, metabolic, regulatory or co-expression networks. These networks are not isolated components within the cell, but strong interconnections exist between the different domains.

Scientific researchers have made use of these networks in order to identify functionally related proteins by exploring the functional context of the proteins, i.e. the interconnected components within the network contributing to a common function. Specialized web resources such as STRING provide facilities for a manual inspection of a single protein's functional context [2].

A number of studies have used entire networks to generate novel insights into the complex organization of biological systems as well as into the relationship between network properties and protein properties such as function or phenotype [3-6]. These studies have used various methods to identify structural units within networks [7], to gather statistical properties of whole networks as well as of individual members of networks [8], or to use networks for specialized tasks such as functional classification of proteins [9]. Since the separation into different types of subnetworks in the cell based on experimental technologies such as protein/protein interactions or transcript data is artificial, other studies have integrated multiple networks to identify functionally related proteins [10,11].

We will introduce the CABiNet (Comprehensive Analysis of Biomolecular Networks) software suite, a generic network analysis system, which is able to span the complete range of aforementioned methods. CABiNet supports functionality for the generation of novel networks and the integration of both, networks and methods.

2 CABiNet – A generic network analysis system

A generic network analysis system needs to provide methods that reach over the range from the exploration of a single protein's functional context to system level analyses. In order to draw a comprehensive picture of the cell's functional organization, these methods must support integration of cellular networks as well as of methods manipulating these networks. We present CABiNet, a system for Comprehensive Analysis of Biomolecular Networks, which is designed to provide all requirements for a generic network analysis system and extends the current range of functionality by providing a semi-automatic network processing pipeline for complex analyses.

2.1 Exploration of a protein's functional context

Even today, with an accelerating number of system level analyses, many scientists are primarily interested in the functional context of a small number of cellular entities to answer "deep drilling" questions. CABiNet aims to provide information both about the local neighborhood of the entities in question as well as to provide results derived from global analyses.

In order to query the functional context of genes and proteins, scientists have to browse through a large number of online resources to get a global view of the interacting partners in the cell, co-expressed genes or functionally related gene products. By choosing a network representation for these associations and providing the functionality to query multiple networks at a time, CABiNet assists the user in finding contextual partners in a large number of independent networks in one single step. Incomplete network information from one method (e.g. due to missing experimental data) may be complemented by information from other sources (see Figure 1A).

By allowing the user to browse the results of network analysis methods, such as clustering, the user can find partners of his entity of interest, which may not be in the immediate neighborhood (see Figure 1B). To go even further, it is possible to search across the whole set of networks and analysis results, combining evidences from multiple networks and methods applied to them. This leads to a more comprehensive picture of how a single entity is embedded into the complex cellular networks (see Figure 1C).

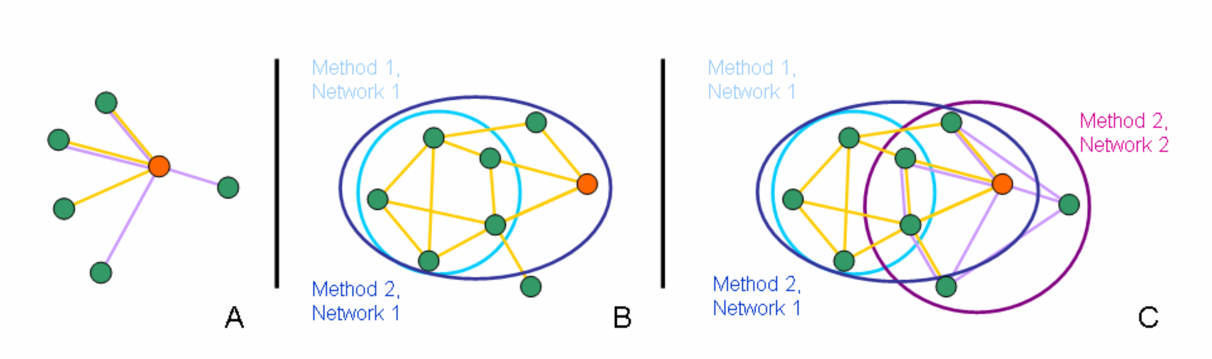

Figure 1 Integration of networks and analysis' results in user queries. When querying for the orange-colored node, results depend on the data set selected. By querying for partners in two networks (A), five neighboring nodes are found, compared to three or four when querying in only a single network (depicted by edge color). When querying for all partners found in for example network clustering methods (B), the result may depend on the method used. Applying method one in this hypothetical example, the node does not belong to any cluster. However, method two adds the orange node to a cluster with six additional members. Figure C provides the most comprehensive picture. Overall, the queried node has eight putative partners worth examining, compared to five found by looking at the two networks independently or a maximum of six found by looking at a single network.

2.2 Integration of networks and methods

To be able to provide users with a wide range of networks, CABiNet allows import of a large variety of input formats. Nodes and edges can be associated with any type of information such as functional annotation or edge weights. In order to map nodes between different networks, the system contains a component capable of resolving established identifiers of molecular entities as well as a component capable of integration of protein networks from two different networks using orthology relationships based on bidirectional best hits in sequence similarity, which is retrieved from SIMAP [12].

The network manipulation methods are comprised of any kind of method, which takes one or more network as input and, based on these networks, either changes the networks or network properties directly or introduces novel network properties. This definition includes methods that change the topological structure of the network as well as algorithms that calculate statistical properties.

In CABiNet, network manipulation methods are categorized into four separate classes, which differ in input parameters and output of the methods. The categories were designed following our experience that almost any kind of network manipulation method fits into one of the given categories. They cover:

- Network conversions, where one input network is transformed.
- Methods calculating statistical properties of the network, leaving the network's topology unchanged.
- Combinatorial methods, where multiple networks are merged.
- Clustering methods, which have the property of identifying substructures within the given network.

This structure allows for an easy addition of new methods into the system.

2.3 Semi-automatic processing pipeline for network analyses

Algorithms applied to network graphs often rely on preliminary manipulations of the input network. For example, an algorithm for functional classification may need a fully annotated integrated network as input, producing a probabilistically annotated output network. In order to generalize this approach, CABiNet supports coupling of any network related algorithms. In the aforementioned case, it is possible to start with multiple graphs in different input formats, which are in the pipeline transformed into the format used internally by CABiNet. In the next step, these networks can be integrated and then annotated, using a Web Service providing methods for functional annotation. This integrated, annotated network is then used as the input for the classification method, which in turn returns another network that could be used in following steps, e.g. for calculation of statistical properties.

In comparison to generic bioinformatics workflow engines, such as for example Taverna [13], CABiNet's processing pipeline solely focuses on biomolecular networks and methods manipulating biomolecular networks, including the representation and storage of results, thereby comprising an efficient network-centric analysis suite.

This network processing pipeline can be accessed from a web user interface. Alternatively, programs can access CABiNet using a Web Service interface.

3 Implementation

The CABiNet framework is programmed in Java, employing business solutions provided by the Java 2 Enterprise Edition (J2EE). Its architecture follows the component-oriented design principle. Component-oriented programming encapsulates functionality into self-contained units with clearly defined interfaces. These components are then used and tied to other components, allowing for great flexibility and re-usability in individual software solutions.

Figure 2 shows the connections between the core components of CABiNet. All network manipulation methods, categorized as described above, are integrated via the standardized external component interfaces in order to be used by the pipeline or the data access and manipulation component. Integration of novel network manipulation methods is straightforward by implementing them as components using the external component interfaces.

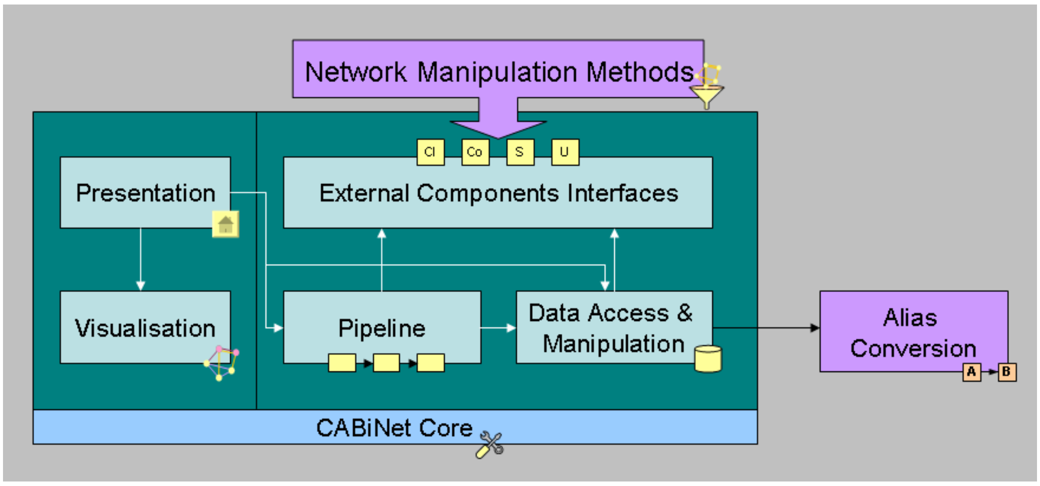
Network Manipulation Methods
Cl
Co
S
U
Presentation
External Components Interfaces
Visualisation
Pipeline
Data Access & Manipulation
Alias Conversion
A
B
CABiNet Core

Figure 2 Component view of CABiNet. Dependencies between components are depicted by arrows. The CABiNet core component (shown at the bottom) provides functionality to all CABiNet related components.

CABiNet already offers a number of methods, spanning all four categories described above. These include various conversion methods responsible for converting different input data into networks, such as PSI-MI files, simple ASCII files or generation of co-expression networks from expression result matrices. Other components provide methods for annotation of networks using any arbitrary Web Service available or for generation of networks of orthologs in order to facilitate cross-genome comparison and to infer networks for non-model organisms from data available for well-studied organisms. The last group of conversion methods is used to functionally classify and statistically interpret functional annotation of biomolecular networks. Various clustering algorithms are grouped in the class of clustering methods. These include MCL clustering [14], clustering based on community structure [15] and a novel algorithm for clustering networks based on local clustering coefficients. Union methods provide methods to integrate different networks, with specialized components for combining networks from two different organisms based on orthology relationships and for uniting two networks in which identifiers from different genome databases are used. Finally, statistic methods include components to derive for example network properties as described by Barabasi and co-workers[16].

4 Sample Workflows

To demonstrate the power of the CABiNet concept, we will introduce three applications in which cellular networks were analyzed using the CABiNet processing pipeline. The versatility of the pipeline is demonstrated by applying the components provided by CABiNet to three very different approaches.

The first study shows how CABiNet can be used to combine biomolecular networks with biological knowledge from genome databases to generate novel insights into the cell's complex structure. In another application example, the pipeline was used to prepare multiple networks for function prediction using a classification algorithm, which was employed at a later stage in the pipeline. Finally, to illustrate CABiNet's capability to work with gene expression data to generate and analyze co-expression networks, gene expression data from time series experiments was used to identify clusters of genes that are co-expressed in the same cell cycle stage.

It has been shown that protein-protein interaction data can be used for identification of functional modules [17-21]. In our first application example, we used a high-confidence protein-protein interaction dataset from yeast [22] and applied the CFinder algorithm, a network clustering method [15], to identify functional modules in the network. In order to assess the quality of the functional modules, functional homogeneity of the proteins within the modules was determined. Functional annotation of proteins from the MIPS Functional Catalogue (FunCat) [23] was used.

In the next step, annotation about a protein's influence on the organism's phenotype was mapped to the proteins in the network. In this study, phenotypic information was restricted to whether a protein is essential for the organism's survival or not. This data was retrieved from the Comprehensive Yeast Genome Database [24]. Functional modules were then examined for the fraction of essential proteins they contain.

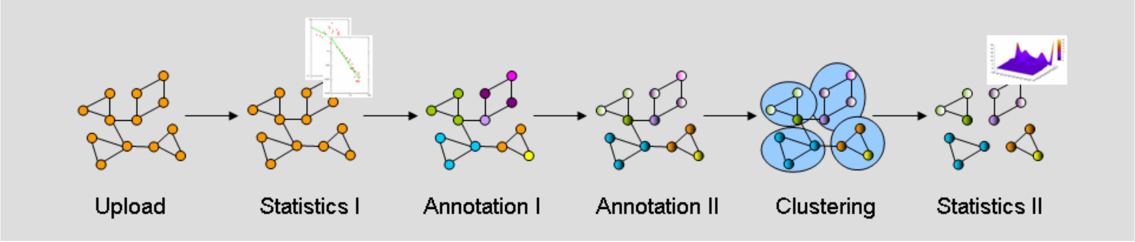

Figure 3 Correlation of phenotypic information and functional modules using CABiNet's processing pipeline. In the first step, a network to be analyzed is uploaded into the system. In order to determine if topological properties support modular structure, network statistics are calculated. In the next two steps, protein annotation is added to the network (functional classification / phenotype). Subsequent clustering leads to functional modules, for which correlation of the annotations is retrieved. The ordering of steps 2-5 in the pipeline has no effect on the results.

To conduct this study, a workflow for CABiNet's processing pipeline, based upon the requirements for the analysis, was constructed (see Figure 3). As mentioned in the requirements, the network to be analyzed needed to be annotated with two independent types of information, once with functional classification of proteins and once with information about the essentiality of the protein. In another step, functional modules were identified using one of the clustering algorithms provided by CABiNet. Finally, the functional annotation was used to determine functional homogeneity of clusters and to illustrate correlations between clusters and phenotypic information.

In the second application (see Figure 4), multiple biomolecular networks of *Neurospora crassa* were processed, integrated and used to infer the cellular function of proteins for which previously no annotation was available.

The processing plan for this approach could be divided into two steps. In the first, the networks necessary for function prediction were uploaded and, if necessary, modified to be used by the classification algorithm. The second subtask combined these networks into one integrated network, which was annotated with protein functions available from the *N.crassa* genome database [25]. Proteins lacking annotation in this network were then assigned a functional category based on a probabilistic classification algorithm available in CABiNet.

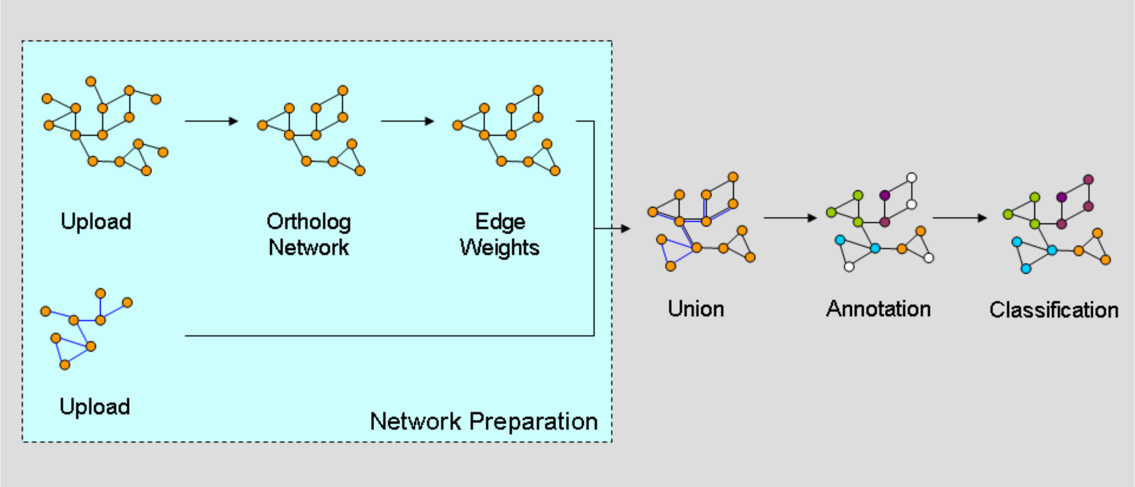

Figure 4 Functional classification using CABiNet's processing pipeline. In this application, CABiNet is used both for preparation of the networks as well as for the classification task. Note that this figure only displays integration of two networks, even though any number of networks may be used.

In the third use case (see Figure 5), gene expression data from *Saccharomyces cerevisiae* was transformed into a network, which was then further analyzed using network clustering techniques in order to obtain clusters of co-expressed genes. These clusters were assessed for their ability to confirm common functional annotation.

The CABiNet network processing pipeline was used to generate a co-expression network from available gene expression data. In an additional step, the network statistics component was used to measure network size and topology. One of the network clustering techniques provided by CABiNet was used to identify clusters of co-expressed genes in the generated network. These clusters were then downloaded from CABiNet and used for comparison of these structures with established results.

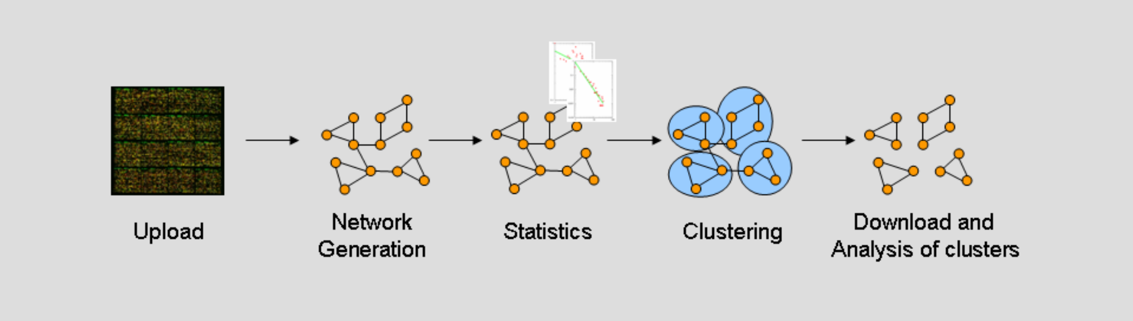

Figure 5 Identification of functional modules from gene expression data. In the first step, uploaded data from gene expression experiments is transformed into a network. In the second step, network statistics are calculated. Finally, the network is clustered and identified clusters are downloaded for further statistical analyses.

5 Discussion

In this paper, we introduced CABiNet as a framework to be used to integrate and concatenate a large number of available and future network analysis methods. The system is based on a stable, seamlessly extendable software platform, which makes it a solid foundation to integrate additional components for network analysis as well as using it as a source for novel applications centered on specific domains of interest.

5.1 Components suitable for integration

The number of network analysis and statistics methods available is already large and continues to expand. In order to answer scientific questions using the CABiNet processing pipeline, relevant methods need to be selected and integrated as CABiNet components using the provided interfaces.

Possible candidates for future inclusion are components for the integration of external networks as well as methods calculating additional network statistic measures and clustering techniques. We present some exemplary components that may be potentially useful for an even more thorough network analysis.

Some topological measures that can be used to describe the network's architecture are available in the network statistics component. However, a large number of additional algorithms are available to characterize complex networks [26]. One or more components capable of deriving these measures might be implemented as CABiNet statistic components, providing the user with a more comprehensive view of a network's organization and allowing for easy comparison of the dependency of the results derived by different methods.

So far, CABiNet uses clustering techniques to detect functional modules, a concept to reduce complexity in biomolecular networks by identifying an organizing principle within them.

Furthermore, it has been shown that certain subgraph patterns, so-called motifs, tend to be significantly overrepresented in these networks [5]. Large efforts have been put into algorithms capable of retrieving these subgraphs in a computationally efficient way [27,28]. These methods could be integrated in CABiNet as network cluster components, storing the motifs as communities of nodes in the network with an attached tag describing the motif pattern, if necessary.

The STRING resource hosts a large number of biomolecular networks [2]. It offers a comprehensive, quality-controlled collection of protein-protein associations for a large number of organisms, from predictions based on genomic context analysis to data derived from mining databases and literature. To facilitate an easy incorporation of these data into the CABiNet system for further analyses using the provided methods, a connection bridge designed as a CABiNet conversion component would be beneficial. By adding this component, inclusion of STRING networks is possible by just specifying the network to be uploaded into CABiNet.

The components described above are only a small selection of putative extensions to the CABiNet system. Since the addition of novel methods is made a feasible task by the framework, these methods can be included as the need for a specific requirement of a network analysis arises.

5.2 Further perspectives

Due to the modularity of the CABiNet system, it can be used to build full-grown self-contained applications based on the framework in conjunction with components designed for the system as well as based solely on individual CABiNet components.

As an example, a system for the semi-automatic annotation of novel genomes may quickly be implemented based on CABiNet's processing pipeline, similar to the use case described earlier. Coupled with a semi-automatic annotation pipeline such as PEDANT[29], which uses protein homology to determine protein function, such a system would be an invaluable tool to enhance function prediction in novel genomes.

As depicted in Figure 6, this system would use the genomic sequence of a newly sequenced organism to generate networks based on genomic context and homology. Additional networks, such as interolog networks [30] or co-expression networks to be included could be uploaded into the system. In the next step, all relevant networks would be merged into one integrated network, which is then annotated using high-confidence annotations from the PEDANT system. In the final step, the classification algorithm could be used to assign functional classes to all proteins without annotation.

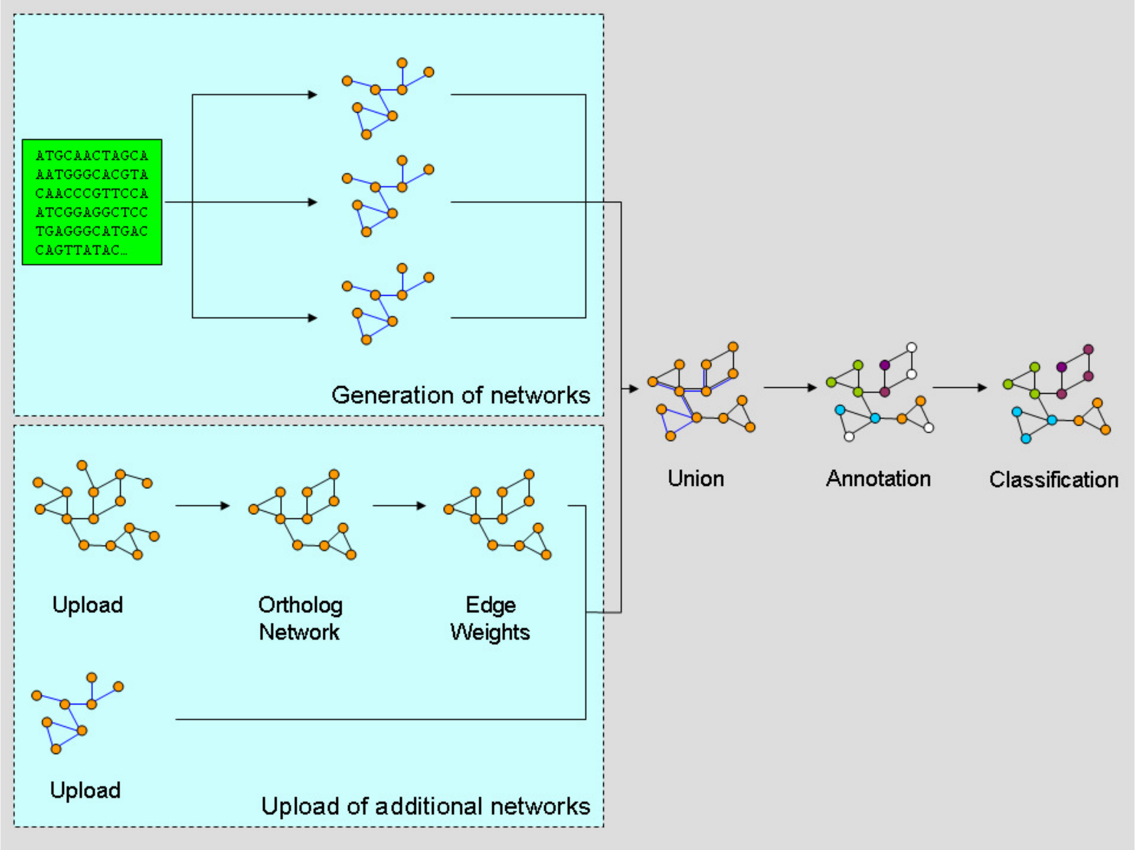

Figure 6 Design of a semi-automatic classification system for novel genomes. This model system consists of three larger parts. One part needs to be capable of generating networks from genomic data. The second part handles uploaded additional networks, while the third part is able to merge the networks into one single network, which is annotated and used for classification.

Additionally, the separation of the presentation layer from the underlying layers providing data access and business logic, allows for individual presentation solutions in specialized applications. As an example, a resource for functional modules in mammals may use all the functionality provided by CABiNet in order to identify and maintain networks in mammals and their associated sets of functional modules and make them available to the public in an individual format. This web application would make use of CABiNet's business methods for retrieval and querying the data of a specific user domain. Since all data is returned in the XML format, presentation of the data is simply a matter of transforming the information into HTML format using XSL stylesheet transformations, made even easier through the availability of ready-to-use XSL stylesheets from the CABiNet system. This leaves only the task of defining page navigation to the developer implementing such a resource.

Therefore, CABiNet's capabilities make it useful as a system for generic network studies as well as a platform for independent novel applications. This makes it an ideal framework to be used to provide novel insights into the complex structure of biomolecular networks. Additionally, it can use these networks as a basis to make predictions about the cellular role of their components.

References

[1] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, From molecular to modular cell biology, Nature, 402 (1999) C47-C52.

[2] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork, STRING: known and predicted protein-protein associations, integrated and transferred across organisms, Nucleic Acids Res., 33 (2005) D433-D437.

[3] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi, Hierarchical organization of modularity in metabolic networks, Science, 297 (2002) 1551-1555.

[4] S. Wuchty, A. L. Barabasi, and M. T. Ferdig, Stable evolutionary signal in a Yeast protein interaction network, BMC. Evol. Biol., 6:8. (2006) 8.

[5] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Network motifs: simple building blocks of complex networks, Science, 298 (2002) 824-827.

[6] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, Lethality and centrality in protein networks, Nature, 411 (2001) 41-42.

[7] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Network motifs in the transcriptional regulation network of Escherichia coli, Nat. Genet., 31 (2002) 64-68.

[8] Z. N. Oltvai and A. L. Barabasi, Systems biology. Life's complexity pyramid, Science, 298 (2002) 763-764.

[9] K. Tsuda, H. Shin, and B. Scholkopf, Fast protein classification with multiple networks, Bioinformatics., 21 Suppl 2 (2005) ii59-ii65.

[10] S. Tornow and H. W. Mewes, Functional modules by relating protein interaction networks and gene expression, Nucleic Acids Res., 31 (2003) 6283-6289.

[11] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir, Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data, Proc. Natl. Acad. Sci. U. S. A, 101 (2004) 2981-2986.

[12] R. Arnold, T. Rattei, P. Tischler, M. D. Truong, V. Stumpflen, and W. Mewes, SIMAP--The similarity matrix of proteins, Bioinformatics., 21 Suppl 2 (2005) ii42-ii46.

[13] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn, Taverna: a tool for building and running workflows of services, Nucleic Acids Res., 34 (2006) W729-W732.

[14] Van Dongen, S. Graph Clustering by Flow Simulation. 2000. University of Utrecht. Ref Type: Thesis/Dissertation

[15] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature, 435 (2005) 814-818.

[16] A. L. Barabasi and Z. N. Oltvai, Network biology: understanding the cell's functional organization, Nat. Rev. Genet., 5 (2004) 101-113.

[17] C. von Mering, E. M. Zdobnov, S. Tsoka, F. D. Ciccarelli, J. B. Pereira-Leal, C. A. Ouzounis, and P. Bork, Genome evolution reveals biochemical networks and functional modules, Proc. Natl. Acad. Sci. U. S. A, 100 (2003) 15428-15433.

[18] B. Snel, P. Bork, and M. A. Huynen, The identification of functional modules from the genomic association of genes, Proc. Natl. Acad. Sci. U. S. A, 99 (2002) 5890-5895.

[19] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis, Detection of functional modules from protein interaction networks, Proteins, 54 (2004) 49-57.

[20] A. W. Rives and T. Galitski, Modular organization of cellular networks, Proc. Natl. Acad. Sci. U. S. A, 100 (2003) 1128-1133.

[21] V. Spirin and L. A. Mirny, Protein complexes and functional modules in molecular networks, Proc. Natl. Acad. Sci. U. S. A, 100 (2003) 12123-12128.

[22] 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, and M. Vidal, Evidence for dynamically organized modularity in the yeast protein-protein interaction network, Nature, 430 (2004) 88-93.

[23] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhaupt, M. Munsterkotter, and H. W. Mewes, The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, Nucleic Acids Res., 32 (2004) 5539-5545.

[24] U. Guldener, M. Munsterkotter, G. Kastenmuller, N. Strack, H. J. van, C. Lemer, J. Richelles, S. J. Wodak, J. Garcia-Martinez, J. E. Perez-Ortin, H. Michael, A. Kaps, E. Talla, B. Dujon, B. Andre, J. L. Souciet, M. J. De, E. Bon, C. Gaillardin, and H. W. Mewes, CYGD: the Comprehensive Yeast Genome Database, Nucleic Acids Res., 33 (2005) D364-D368.

[25] H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, and A. Ruepp, MIPS: analysis and annotation of proteins from whole genomes, Nucleic Acids Res., 32 Database issue (2004) D41-D44.

[26] Costa, L. da F., Rodrigues, F. A., Travieso, G., and Villas Boas, P. R. Characterization of Complex Networks: A Survey of Measurements. arXiv:cond-mat/0505185 v5 . 2006.
Ref Type: Electronic Citation

[27] S. Wernicke and F. Rasche, FANMOD: a tool for fast network motif detection, Bioinformatics., 22 (2006) 1152-1153.

[28] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs, Bioinformatics., 20 (2004) 1746-1758.

[29] D. Frishman, M. Mokrejs, D. Kosykh, G. Kastenmuller, G. Kolesov, I. Zubrzycki, C. Gruber, B. Geier, A. Kaps, K. Albermann, A. Volz, C. Wagner, M. Fellenberg, K. Heumann, and H. W. Mewes, The PEDANT genome database, Nucleic Acids Res., 31 (2003) 207-211.

[30] H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J. D. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein, Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs, Genome Res., 14 (2004) 1107-1118.