From the Book:
Data Analysis and Visualization in Genomics and Proteomics. F. Azuaje, J. Dopazo (Eds). Wiley 2005
Pp. 29-39
ISBN: 0-470-09439-7

3

Data and Predictive Model Integration: An Overview of Key Concepts, Problems and Solutions

Francisco Azuaje, Joaquin Dopazo, Haiying Wang

Abstract

This chapter overviews the combination of different data sources and techniques for improving functional prediction. Key concepts, requirements and approaches are introduced. It discusses two main strategies: a) Integrative data analysis and visualisation approaches with an emphasis on the processing of multiple data types or resources; and b) integrative data analysis and visualisation approaches with an emphasis on the combination of multiple predictive models and analysis techniques. It also illustrates problems in which both methodologies can be successfully applied.

Key words: Integrative data mining, integrative data visualisation, gene expression analysis, protein networks, functional prediction.

3.1 Integrative data analysis and visualisation: Motivation and approaches

The combination of multiple data sources is both a fundamental requirement and goal for developing a large-scale and dynamic view of biological systems. Data originating from multiple levels of complexity and organisation are interrelated to assess their functional predictive abilities. For instance, quantitative relationships between gene expression correlation and protein-protein interaction, gene and protein expression correlation have been studied (Allocco, Kohane, and Butte, 2004). Typical questions addressed by such studies include: Is there a significant connection between highly expressed genes and highly expressed proteins? Is the expression correlation exhibited by a pair of genes significantly associated with the likelihood of finding their products in the same protein complex? These quantitative relationships support the design of prediction models to facilitate functional classification and interpretation. In a post-genomic scenario the possibility of answering functional questions on one-gene-at-a-time bases is being abandoned in favour of more systemic approach in which the accuracy of the individual result is sacrificed at the exchange for a more deep knowledge on how the different parts of the whole system interact among them to play different biological roles. Thus, function is starting to be understood as a more complex concept within systems biology than the naïve adscription of a given activity or role to a single protein.

A massive collection of computational and statistical techniques are available to analyse and visualise different types of "omic" information. The most important computational question is not whether there are options for a particular problem. Rather, bioinformaticians are becoming more concerned about questions such as: *how to combine different techniques? When? Why?*

The combination of multiple prediction models is fundamental to address limitations and constraints exhibited by individual approaches. Moreover, their integration may improve the accuracy, reliability and understandability of prediction tasks under different experimental and statistical assumptions and conditions. For example, it has been demonstrated that the combination of multiple, diverse classification models may significantly outperform the prediction outcomes obtained from the application of individual classifiers (Kittle, Hatef, Duin, and Matas, 1998). Thus, model diversity is a crucial factor to achieve multiple-views of the same problem, reduce bias and improve the coverage of the prediction space. Diversity may be obtained not only through the application of multiple models, but also though the implementation of different methods for selecting data, features and prediction outcomes.

In general two major computational categories of integrative data analysis and visualisation approaches may be identified: a) Those approaches that place an emphasis on the processing of multiple data types; and b) those approaches that rely on the combination of multiple predictive models and analysis techniques. The first approach may of course apply multiple predictive computational models, but its main goal is to combine different types of biological data sets in order to improve a prediction task or to achieve a more complete, dynamic view of a biological problem. An example of this type of approach is the combination of expression, cellular localization and protein interaction data for the prediction of protein complex membership. Although the second approach may (or may not) process different types of data, its main objective is to implement different statistical and/or machine learning models to improve predictive quality. One example is the combination of several clustering algorithms, including neural networks, to improve accuracy and coverage in the functional characterisation of genes based on microarray data.

This chapter discusses these two main data analysis and visualisation problems by providing an overview of recent key investigations and applications for functional genomics. It also illustrates problems in which both methodologies can be successfully applied.

3.2 Integrating informational views and complexity for understanding function

The organisational modules of the cell may be divided into several types of "*omic*" information. For example, the transcriptome refers to the set of information transcribed from coding sequences, which is defined by their expression patterns. The interactome specifies the existing interactions between molecules in the cell, including protein-protein and protein-DNA interactions. The reader is referred to (Ge, Walhout, and Vidal, 2003) for a discussion on the classification of "omic" approaches.

Information originating from each "omic" approach may be incomplete, incorrect or irrelevant. Their predictive quality and usefulness may be significantly compromised by the presence of several false negatives and false positives. Each data source offers a different, partial view of the functional roles of genes and proteins. But also they may generate overlapping views of the same problem. Therefore, their integration may provide the basis for more effective and meaningful functional predictors. Moreover, it may support the generation and validation of new hypotheses. For instance, if *method A* suggests that gene product *X* interacts with gene product *Y*, it would be then important to apply other methods to assess the relevance or validity of this interaction. Phenotypic information describing the essentiality of these genes together with their expression patterns may aid in the identification of their participation in common biological pathways or related functions. Thus, these putative roles may reflect the relevance of this interaction.

An integrative prediction process aims to exploit the existing quantitative relationships between different "omic" data sets. These relationships may indicate the type of constraints and integration mechanisms that need to be defined. Thus, for instance, an important problem is to investigate how different data sets are statistically correlated. In some applications is important to assess the significance of such relationships with respect to relationships detected from random data sets. Advances in this area include techniques to describe how gene expression correlation and interactome data are interrelated in *S. cerevisiae*. Several correlation measures, such as the *Pearson coefficient* and the *cosine distance*, may be used. A typical strategy consists of depicting the distribution of expression correlation values for interactome data sets, which may be compared with the distribution obtained from random protein pairs (Ge *et al.*, 2003). These comparisons indicate, for example, that interacting proteins are more likely to be encoded by genes strongly correlated by their expression profiles (Jansen *et al.* 2003). Another technique consists of plotting the likelihood of finding two proteins in the same protein complex as a function of their expression correlation coefficients (Jansen, Greenbaum, and Gerstein, 2002). The validity of this methodology for detecting transcriptome-interactome relationships in multi-cellular organisms requires further investigation. For instance, it has been suggested that these relationships can be observed in *C. elegans* at least for particular types of tissue (Walhout *et al.*, 2002).

This data visualisation procedure may be easily extended to estimate other functional properties, such as the likelihood of finding pairs of genes regulated by a common transcription factor on the basis of their gene expression correlation. It has been shown that pairs of genes with significantly correlated expression patterns are much likelier to be bound by a common transcription factor, in comparison to those pairs exhibiting weaker expression correlations (Allocco *et al.*, 2004).

Interrelationships between interactome and phenome, transcriptome and translatome, and transcriptome and phenome have also been studied (Ge *et al.*, 2003). Such associations may motivate different interpretations, which sometimes may be specific to particular organisms or functional roles. But which may be reconciled and integrated to formulate hypotheses or to support the development of more effective prediction models (Ge *et al.*, 2003). Figure 3.1 illustrates typical plots for visualising potential significant relationships between different ‘omic’ properties.

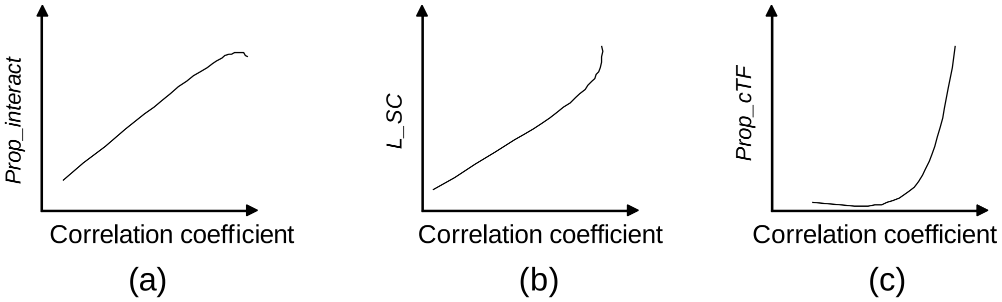

Figure 3.1. Typical plots used to identify relevant relationships between different “omic” data sets (hypothetical examples). (a) Displaying relationships between the proportion of interacting proteins (*Prop_interact*) versus their correlation coefficients. (b) The likelihood of finding two proteins in the same complex (*L_SC*) versus their correlation values. (c) The proportion of pairs of genes bound by a common transcription factor (*Prop_cTF*) versus their correlation.

Once potential relationships have been identified, models may be built to combine evidence or prediction outcomes derived from different data sources. Several machine learning methods, such as decision trees and neural networks, may be applied to implement this task. For instance, integrative models based on Bayesian networks have been applied to predict protein-protein interactions in yeast. One recent advance (Jansen *et al.*, 2003) reported the integration of different types of experimental interaction data, functional annotations, mRNA expression and essentiality data to improve the identification of protein-protein interactions. One important advantage shown by probabilistic frameworks is that they provide an assessment of the predictive relevance and reliability of each integrated source. They are useful to deal with different types of data and missing values. Moreover, relationships between sources are expressed in terms of conditional probabilities, which in many applications facilitate the interpretation of results. One limitation is that these models often require the user to make strong assumptions about the independence of the information sources, which may not be easy to justify or accurate to generate reliable predictions.

Integrative data analysis approaches are also fundamental tools for refining or adapting other systemic models such metabolic networks (Ideker *et al.*, 2001). In this case different types of data, such as mRNA expression, protein expression and physical interaction data may be used to measure responses to systematic perturbations. Data clustering techniques and correlation visualisation tools (including those discussed above) may be applied to summarise these responses and their associations with functional roles or processes.

One important problem that requires further research is the development of methods to visualise not only different information sources, but also multiple analysis outcomes. These techniques should support both interactive and iterative tasks. A key limitation, which was discussed in the Chapter 1, is that the areas of data analysis (or data mining) and visualisation have traditionally evolved as separate disciplines. Typical information visualisation tools have been designed to process single data sources. Moreover, they have put emphasis on the problem of displaying final analysis outcomes, without providing more hierarchical, multi-resolution views of prediction processes. Thus, an integrative data visualisation approach is necessary not only to complement integrative data analyses, but also to make them more meaningful.

Information visualisation platforms currently available allow researchers to merge multiple data sources to highlight relevant relationships, such as those represented in regulatory networks (Baker, Carpendale, Prusinkiewicz, and Surette, 2002). Regulatory networks may be, for instance, displayed together with other types of information such as gene expression correlation and interaction information. Different experimental methods or relationships may be represented by using colour coding schemes associated with the nodes and edges in the network.

Integrative visualisation tools should provide multiple graphical and analytical views of other organisational levels or “omic” sources, including pathways and functional annotations. The *VisAnt* platform is one of such options (Hu, Mellor, Wu, and DeLisi, 2004), in which metabolic data, gene homology, annotations and cross-referencing information of genes and proteins are integrated. One important challenge for this type of research is to support a flexible, open and integrated display of heterogeneous information sources and analysis outcomes. In this direction, ambitious projects of genome browsers such as the Ensembl (Birney et al., 2004) with tools for such as EnsMart (Kasprzyk et al., 2004) allows easy integration of different types of information in a genomic context and with the possibility of cross-genome comparisons. Similar tools are the NCBI’s Map viewer or the UCSC genome browser. All these tools incorporate genomic information and make it available through friendly web browsers.

A fundamental condition to achieve an integrative data analysis and visualisation paradigm is the ability to integrate diverse outcomes originating from the application of multiple prediction models.

3.3 Integrating data analysis techniques for supporting functional analysis

One important characteristic exhibited by the models introduced above is that they combine multiple data sources by mainly applying only one type of prediction model, such as a single classification technique. An alternative integrative prediction approach may also take advantage of the diversity of available prediction models and techniques. It has been demonstrated that different techniques can unveil various aspects of different types of data such as gene expression data (Leung and Cavalieri, 2003). The combination of diverse models can overcome the dependency on problem- or technique-specific solutions.

One such integrative approach is known as *Multisource Association of Genes by Integration of Clusters*, which was proposed by Troyanskaya and co-workers (Troyanskaya, Dolinski, Owen, Altman, and Botstein, 2003). It applies probabilistic reasoning and unsupervised learning to integrate different types of large-scale data for functional prediction. The system has been tested on *S. cerevisiae* by combining multiple classification techniques based on microarray, physical and genetic interactions and transcription factor binding sites data. An assessment of functional prediction relevance in yeast has been performed by processing Gene Ontology annotations derived from the *S. cerevisiae* Genome Database. The inputs to the integrative probabilistic prediction framework may consist of clustering-driven predictions based on gene expression correlation and other functional relationships between pairs of gene products. This framework allows, for instance, the combination of classification outcomes generated by several clustering techniques such as *k*-means, self-organising maps and hierarchical clustering. The system estimates the probability that a pair of gene products is functionally interrelated. Such a relationship is defined by their involvement in the same biological process, as defined by the Gene Ontology. This approach clearly demonstrates how an integrative approach may outperform single-source prediction techniques, such as models based solely on microarray data. Moreover, it highlights the advantages of combining multiple classification methods. Troyanskaya further discusses this integrative framework and its applications in Chapter 11.

Other authors, such as Wu *et al.* (2002), have showed the importance of applying multiple clustering methods to discover relevant biological patterns from gene expression data. This type of models aims to integrate classification outcomes originating from several clustering methods such as: Hierarchical clustering, *k*-means, and self-organising maps. One important assumption is that these methods may produce partially overlapping expression clusters. Multiple partitions may be obtained by running different clustering algorithms using several learning parameters or numbers of clusters. Without going into details, a functional class prediction derived from a clustering experiment may be associated with a probability value, *P*. It estimates the possibility that a cluster of genes was obtained by chance and allows assigning a gene to multiple functional categories. Thus, integrative predictions are made on the basis of the minimum *P*-value exhibited by a category in a cluster. The computational predictions and experimental validation performed by Wu *et al.* further demonstrate the importance of integrating several machine learning and statistical methods to improve biological function predictions based on a single data source. One key advantage of combining multiple clustering-based prediction outcomes is that it allows the association of multiple, reliable functional predictions to a gene product based on a probabilistic framework. Clusters may be automatically linked to significant functional categories by processing a reference knowledge base, such as the Gene Ontology. The implementation of tools for automatically annotating clusters is a fundamental problem to achieve integrative data analysis goals. In Chapter 7, Al-Shahrour and Dopazo will discuss the problem of assigning significant functional classes to gene clusters based on Gene Ontology annotations. Figure 3.2 summarises basic tasks required in a clustering-driven integrative framework for predicting functional classes.

In Chapter 10 Sheng and co-workers review several clustering techniques and methods for assessing the statistical quality of clusters. Different statistical methods may be combined to support the evaluation of clusters in terms of their significance, consistency and validity. This is a problem that deserves more attention and investigation in order to improve the design and interpretation of functional genomics studies, especially those analyses based on gene expression clusters. For instance, the application of *null hypothesis tests*, *internal* and *external validity indices* may be applied to select relevant, significant partitions and clusters. The estimation of the "correct number of clusters" represented in a dataset is a complex task, which may strongly influence the products of a predictive analysis process. These tests may be used for: a) providing evidence against the hypothesis: "there are no clusters in the data" (null hypothesis tests); b) for finding the optimal partition on the basis of several inter- and intra-cluster distances (internal validity indices); or c) for assessing the agreement between an experimental partition and a reference partition (external indices). The experimental partition is the partition under study, while the reference dataset may be a partition with *a priori* known cluster structure. Bolshakova and Azuaje (2003) have proposed strategies to integrate the outcomes originating from multiple cluster validity indicators, which may be used to generate more reliable and robust predictions about the correct number of clusters.

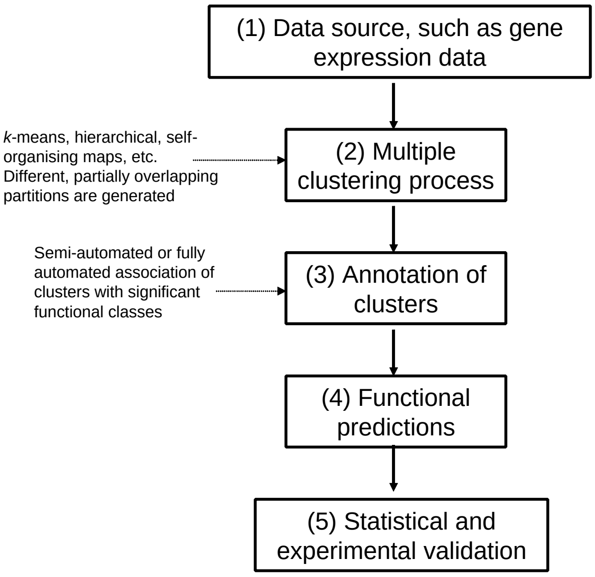
(1) Data source, such as gene expression data
k-means, hierarchical, self-organising maps, etc. Different, partially overlapping partitions are generated
(2) Multiple clustering process
Semi-automated or fully automated association of clusters with significant functional classes
(3) Annotation of clusters
(4) Functional predictions
(5) Statistical and experimental validation

Figure 3.2. Clustering-based integrative prediction framework: Basic tasks and tools. Different, partially overlapping partitions are generated by implementing different clustering techniques, based on different learning parameters and numbers of clusters. Probabilistic assessment about the significance of clusters in relation to functional categories is required for automatically labeling clusters and assigning classes to genes.

3.4 Final remarks

The goal of integrative data analysis and visualisation is not only to increase the accuracy and sensitivity of functional prediction tasks, but also to achieve better insights into the problems under consideration. Even when this type of approaches has become of great importance in genomics and proteomics, the problem of combining a wide variety of information to form a coherent and consistent picture of functional prediction problems has lagged. Moreover, current advances combine different types of data relying on the application of a single prediction model (Zhang, Wong, King, and Roth, 2004), which often are based on strong assumptions about the statistical independence or distribution of the data under study (Jansen et al. 2003). To fully exploit integrative data and visualisation there is a need to process data derived from different sources. Similarly, it is fundamental to combine diverse predictive views originating from multiple classifiers or prediction models. Furthermore, it is crucial to continue studying relationships between apparently unrelated data, which may provide the basis for novel prediction information sources and models to be integrated.

Sections 3.2 and 3.3 overviewed two key strategies to perform integrative data analysis and visualisation in functional genomics. Within such an integrative framework it is also possible to define problems, methods and applications according to: a) the type of data integration, and b) the level in which predictive model integration is achieved.

According to the type of data integration, integrative approaches can be categorised as follows.

Redundant information integration approaches: These approaches process information provided from a group of sources that represent the same type of functional data, e.g: expression data, but with a different degree of accuracy or confidentiality. Applications may be based on the integration of replicated sources that measure similar properties, but which may be noisy, inaccurate or subject to statistical variations (Edwards *et al.,* 2002). They generally aim to reduce the overall uncertainty and increase the predictive accuracy.

Complementary information integration approaches: These approaches integrate information from sources that represent different variables or properties of the prediction problem under consideration. Complementary information integration aims at combining partial, incomplete and noisy information to get a global picture of the prediction problem domain. One typical example is the combination of expression and interaction data sets to predict complex membership. Multiple sources provide information that may be not perceived by using individual experimental methods.

According to the level in which information integration is performed, problems and applications can be categorised as follows.

Integration at the level of input representation: Information provided from the sources is fused before performing prediction or classification tasks. This process may be implemented by integrating in a unique input feature vector the attribute values that represent the different variables under study. For instance, Zhang *et al.* (2004) grouped several gene- and protein-pair properties into a single binary feature representation to predict co-complexed pairs in *S. cereivisiae* based on decision trees.

Integration at the level of feature pre-processing: In this case the product of different feature filtering or selection procedures applied to an information source is combined before performing a classification task. Based on a combination of several feature selection schemes, including *signal-to-noise ratio* and an *evolving classification function* technique, Goh *et al.* have recently introduced a hybrid feature selection method to improve classification of gene expression data (Goh, Song, and Kasabov, 2004). This study highlighted the advantages of a hybrid, integrative method for gene selection.

Integration at the level of classification: Information provided from different sources or prediction models is processed independently, their prediction outcomes are generated, and then integrated in order to make a final prediction about the functional problem under consideration. One example from this category is the integration of serial and parallel competitive classifiers such as ensembles of neural networks and decision trees (Tan and Gilbert, 2003; Hu and Yoo, 2004).

The application of integrative data analyses at the pre-processing and classification levels based on different types of functional data deserves further investigation. It may offer powerful tools not only to improve predictive quality (accuracy and coverage), but also to support the generation of more comprehensive studies at a systems level.

As a final caveat, it is important to remark that, while on one hand the overabundance of data can fuel our understanding on the living systems, on the other hand, the possibility of observing expurious associations between genes and functional properties due to pure chance cannot be neglected. It is necessary then to establish a rigorous framework for the analysis of data at gnomic scale.

References

Allocco, D. J., Kohane, I. S., and Butte, A. J. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics, 5(18).

Baker, C. A. H., Carpendale, M. S. T., Prusinkiewicz, P., and Surette, M. G. (2002) GeneVis: visualisation tools for genetic regulatory network dynamics. In *Proceedings of 13th IEEE Visualisation 2002 conference* (pp. 243-250). Boston, MA.

Bolshakova, N., and Azuaje, F. (2003) Cluster validation techniques for genome expression data. *Signal Processing*, 83(4), 825-833.

Birney, E. et al. (2004) Ensembl Nucleic Acids Research 32: D468-D470

Edward, A. M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J., and Gerstein, M. (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet., 18, 529-536.

Ge, H., Walhout, A. J. M., and Vidal, M. (2003) Integrating 'omic' information: a bridge between genomics and system biology. *Trends in Genetics*, 19(10), 551-560.

Goh, L., Song, Q., and Kasabov, N. (2004) A novel feature selection method to improve classification of gene expression data. In *Proceedings of the second conference on Asia-Pacific bioinformatics: vol. 29* (pp. 161-166), Dunedin, New Zealand.

Hu, Z., Mellor, J., Wu, J., and DeLisi, C. (2004) VisANT: an online visualisation and analysis tool for biological interaction data. *BMC Bioinformatics*, 5(17).

Hu, X., and Yoo, I. (2004) Cluster ensemble and its applications in gene expression analysis. In *Proceedings of the second conference on Asia-Pacific bioinformatics: vol. 29* (pp. 297-302), Dunedin, New Zealand.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(4), 929-934.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Research*, 12(1), 37-46.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science,* 302 (17), 449-453.

Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. 2004. EnsMart—A generic system for fast and flexible access to biological data. *Genome Res.* 2004 **14:** 160-169.

Kittle, J., Hatef., M., Duin, R. P. W., and Matas, J. (1998) On Combining Classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(3), 226-239.

Leung, Y. F. and Cavaloeri, D. (2003) Fundamentals of cDNA microarray data analysis. *Trends Genet.* 19(11), 649-659.

Tan, A. C. and Gilbert, D. (2003) Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2(3 Suppl), S75-S83.

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(14), 8348-8353.

Walhout, A. J., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P., Ge, H., Lee, H., Doucette-Stamm, L., Gunsalus, K. C., Schetter, A. J., Morton, D. G., Kemphues, K. J., Reinke, V., Kim, S. K., Piano, F., and Vidal, M. (2002) Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.*, 12, 1952-1958.

Wu, L. F., Hughes, T. R., Davierwala, A. P., Robinson, M. D., Stoughton, R., and Altschuler, S. J. (2002) Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nat. Genet.*, 31, 255-265.

Zhang, L. V., Wong, S. L., King, O. D., and Roth, F. P. (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration**.** *BMC Bioinformatics*, 5(38).