UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

SCUOLA DI DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE
INDIRIZZO BIOINGEGNERIA
CICLO XXIV

BIOMARKER LISTS STABILITY IN GENOMIC STUDIES: ANALYSIS AND IMPROVEMENT BY PRIOR BIOLOGICAL KNOWLEDGE INTEGRATION INTO THE LEARNING PROCESS

Direttore della Scuola: Ch.mo Prof. Matteo Bertocco

Coordinatore d'indirizzo: Ch.mo Prof. Claudio Cobelli

Supervisore: Ch.ma Dott.ssa Barbara Di Camillo

Dottoranda: TIZIANA SANAVIA

Contents

List of Tables

List of Figures

Abstract

The analysis of high-throughput sequencing, microarray and mass spectrometry data has been demonstrated extremely helpful for the identification of those genes and proteins, called *biomarkers*, helpful for answering to both diagnostic/prognostic and functional questions. In this context, robustness of the results is critical both to understand the biological mechanisms underlying diseases and to gain sufficient reliability for clinical/pharmaceutical applications. Recently, different studies have proved that the lists of identified biomarkers are poorly reproducible, making the validation of biomarkers as robust predictors of a disease a still open issue. The reasons of these differences are referable to both data dimensions (few subjects with respect to the number of features) and heterogeneity of complex diseases, characterized by alterations of multiple regulatory pathways and of the interplay between different genes and the environment. Typically in an experimental design, data to analyze come from different subjects and different phenotypes (*e.g.* normal and pathological). The most widely used methodologies for the identification of significant genes related to a disease from microarray data are based on computing differential gene expression between different phenotypes by univariate statistical tests. Such approach provides information on the effect of specific genes as independent features, whereas it is now recognized that the interplay among weakly up/down regulated genes, although not significantly differentially expressed, might be extremely important to characterize a disease status. Machine learning algorithms are, in principle, able to identify multivariate nonlinear combinations of features and have thus the possibility to select a more complete set of experimentally relevant features. In this context, supervised classification methods are often used to select biomarkers, and different methods, like discriminant analysis, random forests and support vector machines among others, have been used, especially in cancer studies. Although high accuracy is often achieved in classification approaches, the reproducibility of biomarker lists still remains an open issue, since many possible sets of biological features (*i.e.* genes or proteins) can be considered equally relevant in terms of prediction, thus it is

in principle possible to have a lack of stability even by achieving the best accuracy. This thesis represents a study of several computational aspects related to biomarker discovery in genomic studies: from the classification and feature selection strategies to the type and the reliability of the biological information used, proposing new approaches able to cope with the problem of the reproducibility of biomarker lists. The study has highlighted that, although reasonable and comparable classification accuracy can be achieved by different methods, further developments are necessary to achieve robust biomarker lists stability, because of the high number of features and the high correlation among them.
In particular, this thesis proposes two different approaches to improve biomarker lists stability by using prior information related to biological interplay and functional correlation among the analyzed features. Both approaches were able to improve biomarker selection. The first approach, using prior information to divide the application of the method into different subproblems, improves results interpretability and offers an alternative way to assess lists reproducibility. The second, integrating prior information in the kernel function of the learning algorithm, improves lists stability.
Finally, the interpretability of results is strongly affected by the quality of the biological information available and the analysis of the heterogeneities performed in the Gene Ontology database has revealed the importance of providing new methods able to verify the reliability of the biological properties which are assigned to a specific feature, discriminating missing or less specific information from possible inconsistencies among the annotations.
These aspects will be more and more deepened in the future, as the new sequencing technologies will monitor an increasing number of features and the number of functional annotations from genomic databases will considerably grow in the next years.

L'analisi di dati high-throughput basata sull'utilizzo di tecnologie di sequencing, microarray e spettrometria di massa si è dimostrata estremamente utile per l'identificazione di quei geni e proteine, chiamati *biomarcatori*, utili per rispondere a quesiti sia di tipo diagnostico/prognostico che funzionale. In tale contesto, la stabilità dei risultati è cruciale sia per capire i meccanismi biologici che caratterizzano le malattie sia per ottenere una sufficiente affidabilità per applicazioni in campo clinico/farmaceutico. Recentemente, diversi studi hanno dimostrato che le liste di biomarcatori identificati sono scarsamente riproducibili, rendendo la validazione di tali biomarcatori come indicatori stabili di una malattia un problema

ancora aperto. Le ragioni di queste differenze sono imputabili sia alla dimensione dei dataset (pochi soggetti rispetto al numero di variabili) sia all'eterogeneità di malattie complesse, caratterizzate da alterazioni di più pathway di regolazione e delle interazioni tra diversi geni e l'ambiente.
Tipicamente in un disegno sperimentale, i dati da analizzare provengono da diversi soggetti e diversi fenotipi (*e.g.* normali e patologici). Le metodologie maggiormente utilizzate per l'identificazione di geni legati ad una malattia si basano sull'analisi differenziale dell'espressione genica tra i diversi fenotipi usando test statistici univariati. Tale approccio fornisce le informazioni sull'effetto di specifici geni considerati come variabili indipendenti tra loro, mentre è ormai noto che l'interazione tra geni debolmente up/down regolati, sebbene non differenzialmente espressi, potrebbe rivelarsi estremamente importante per caratterizzare lo stato di una malattia. Gli algoritmi di machine learning sono, in linea di principio, capaci di identificare combinazioni non lineari delle variabili e hanno quindi la possibilità di selezionare un insieme più dettagliato di geni che sono sperimentalmente rilevanti. In tale contesto, i metodi di classificazione supervisionata vengono spesso utilizzati per selezionare i biomarcatori, e diversi approcci, quali discriminant analysis, random forests e support vector machines tra altri, sono stati utilizzati, soprattutto in studi oncologici. Sebbene con tali approcci di classificazione si ottenga un alto livello di accuratezza di predizione, la riproducibilità delle liste di biomarcatori rimane ancora una questione aperta, dato che esistono molteplici set di variabili biologiche (*i.e.* geni o proteine) che possono essere considerati ugualmente rilevanti in termini di predizione. Quindi in teoria è possibile avere un'insufficiente stabilità anche raggiungendo il massimo livello di accuratezza.
Questa tesi rappresenta uno studio su diversi aspetti computazionali legati all'identificazione di biomarcatori in genomica: dalle strategie di classificazione e di feature selection adottate alla tipologia e affidabilità dell'informazione biologica utilizzata, proponendo nuovi approcci in grado di affrontare il problema della riproducibilità delle liste di biomarcatori. Tale studio ha evidenziato che sebbene un'accettabile e comparabile accuratezza nella predizione può essere ottenuta attraverso diversi metodi, ulteriori sviluppi sono necessari per raggiungere una robusta stabilità nelle liste di biomarcatori, a causa dell'alto numero di variabili e dell'alto livello di correlazione tra loro.
In particolare, questa tesi propone due diversi approcci per migliorare la stabilità delle liste di biomarcatori usando l'informazione a priori legata alle interazioni biologiche e alla correlazione funzionale tra le features analizzate. Entrambi gli approcci sono stati in grado di migliorare la selezione di biomarcatori.

Il primo approccio, usando l'informazione a priori per dividere l'applicazione del metodo in diversi sottoproblemi, migliora l'interpretabilità dei risultati e offre un modo alternativo per verificare la riproducibilità delle liste. Il secondo, integrando l'informazione a priori in una funzione kernel dell'algoritmo di learning, migliora la stabilità delle liste.
Infine, l'interpretabilità dei risultati è fortemente influenzata dalla qualità dell'informazione biologica disponibile e l'analisi delle eterogeneità delle annotazioni effettuata sul database Gene Ontology rivela l'importanza di fornire nuovi metodi in grado di verificare l'attendibilità delle proprietà biologiche che vengono assegnate ad una specifica variabile, distinguendo la mancanza o la minore specificità di informazione da possibili inconsistenze tra le annotazioni.
Questi aspetti verranno sempre più approfonditi in futuro, dato che le nuove tecnologie di sequencing monitoreranno un maggior numero di variabili e il numero di annotazioni funzionali derivanti dai database genomici crescerà considerevolmente nei prossimi anni.

Chapter 1

Introduction

1.1 Biomarker discovery and stability in genomic studies

Transcriptome analysis performed with high-throughput microarrays [1] experienced a huge diffusion and profoundly changed the approach to the study of complex diseases, becoming a commonly used tool in biological and medical research due to its ability to simultaneously profile the expression of thousands of genes. In a typical experimental design, data come from different subjects and phenotypes. The analysis of these data has proven extremely useful for the identification of genes and proteins for the development of new physiological hypotheses useful for answering diagnostic, prognostic and functional questions. Numerous studies have for example investigated the so-called molecular signatures, *i.e.* predictive models based on the expression of a small number of genes in order to guide the need for adjuvant therapy [2, 3, 4]. Different methods have been developed to address this issue. Widely used methodologies are based on computing differential gene expression by univariate methods, which calculate a statistic (often a t-statistic) for each gene, measuring differential expression on different experimental conditions. A p-value is usually generated for each gene, based on the statistic, via permutation or a parametric distribution. To account for the thousands of comparisons performed, procedures controlling the false discovery rate (FDR) [5] are applied. Genes that survive the correction for multiple comparisons are then considered differentially expressed while genes that fail to meet the criterion for significance are non-differentially expressed. Once obtain the list of differentially expressed genes, it is the responsibility of the biomedical researcher to draw further conclusions. Such an approach provides information on the effects

of specific genes as individual features, whereas it is now widely recognized that the interplay between weakly up/down regulated genes, although not significantly differentially expressed, might be extremely important and can act on regulatory pathways as significantly differentially expressed genes do. In particular, highly differentially expressed genes tend to be "downstream" genes. Many upstream proteins, such as transcription factors and other regulatory proteins, may only show very moderate changes, especially in contrast to high abundance proteins expressed at the end of the biological cascade. If attention is restricted to only the most highly differentially expressed genes, upstream effects are likely to be missed, despite the crucial role they play, acting as activators.
Other methods, derived from machine learning theory, are characterized by inductive algorithms, *i.e.* algorithms that learn from examples on a given domain, providing a model able to classify new biological samples by identifying multivariate nonlinear features. Beside the classification problem, in order to effectively select those features able to explain alterations characterizing the disease, feature selection algorithms give the possibility to select a set of experimentally relevant gene features.
From a methodological point, in order to both achieve a high predictive performance and effectively select relevant predictors from microarray gene expression data, many statistical learning methods, combined with different feature selection strategies, have been adapted or developed in order to deal with high-dimensional data. Several classifiers such as linear and quadratic discriminant analysis [6, 7] and Support Vector Machines [8] can be opportunely associated to a feature selection phase. These approaches can belong to three main categories: filters, wrappers and embedded. Filter methods rank all variables in terms of relevance, as measured by a score that depends on the classification method, without affecting the learning process. Wrapper methods attempt to select jointly sets of features with good predictive power. Since testing all combinations of features is computationally impossible, wrapper methods usually perform a greedy search in the space of sets of features, *e.g.* Support Vector Machines with Recursive Feature Selection (see Chapter 2). Embedded methods are learning algorithms that perform feature selection in the process of training: the search for an optimal subset of features is built into the classifier construction, and it can be seen as a search in the combined space of feature subsets and hypotheses (see Chapter 3). These methods have attracted strong research interest, in particular for the biomarker discovery task [9]. A biomarker may be defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biologi-

cal processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [10]. The discovery of biomarkers from high-throughput data is typically modeled by selecting the most discriminating features (usually genes or proteins) for classification (*e.g.* discriminating healthy versus diseased subjects) [11].
However, the massive analysis of these high-throughput data carried out in different laboratories and research centers has revealed how difficult is to reproduce the results when the experiments are repeated. In other words, the list of key features for the identification of a pathology only partially overlap when the experimental protocol is replicated in different laboratories or, sometimes, in the same laboratory [12]. The reasons for these differences are, at this moment, source of important scientific discussions and seem to be imputable to different causes:

1. Datasets often include small numbers of subjects (some tens) with respect to the number of variables (tens of thousands of genomic probes in human) [13, 14];

2. The most complex pathologies, such as cancer, are heterogeneous and multicausal, as a result of the alteration of multiple regulatory pathways and of the interplay between different genes and the environment, rather than imputable to a single dysfunctional gene like in monogenic diseases [15];

3. Different laboratories may use different or poorly reproducible experimental protocols and analysis pipelines to process biological samples and data [16, 17].

The analysis of the stability of biomarker selection techniques is only a topic of recent interest [18], and it has not yet considered into the mainstream methodology for biomarker discovery [19]. Although many feature selection and classification algorithms have been proposed, they do not necessarily identify the same candidate feature subsets if the biomarker discovery procedure is repeated [20]. Even for the same data, one may find many different subsets of features (either from the same feature selection method or from different feature selection methods) that can achieve the same or similar predictive accuracy [21, 22, 23]. An unfortunate consequence of this lack of stability is that the biological interpretation of possible functions and pathways underlying the biomarker list is difficult a posteriori.

1.2 Advances in biomarker discovery with gene expression data

Referring to the first two causes of instability highlighted in the previous section, different new approaches for biomarker discovery have been developed recently. In situations where the number of features (genes or proteins) is intrinsically higher than the number of available patients as in high-throughput data, standard methods from statistical learning fail to deal correctly with the so-called "curse of dimensionality". Dimension reduction methods, such as principal component analysis or partial least squares (see [24] for an overview), try to overcome this problem by merging the many available covariates into some few by using linear combinations of original input variables for classification task [25, 26]. Although they may also lead to satisfactory classification performance, biomedical implications of the classifiers are usually not obvious, since all input variables are used in construction of the super variables which are poorly interpretable in a biological sense. Moreover, classic feature selection methods aim at selecting a minimum subset of features to construct a classifier of the best predictive accuracy, often ignoring "stability" in the algorithm design.

To address this issue, ensemble feature selection methods are the first attempt to incorporate stability considerations into the algorithm design step. Ensemble learning methods combine multiple learned models under the assumption that "two (or more) heads are better than one". Many feature selection methods are known to be sensitive to small perturbations of the training data, resulting in unstable signatures. In order to "stabilize" variable selection, several works have proposed to use ensemble feature selection on bootstrap samples: the variable selection method is run on several random subsamples of the training data, and the different lists of selected variables are merged into a hopefully more stable subset [27].

However, most biomarker discovery applications typically assume that all features are equally relevant before the selection procedure. An alternative solution for the stability problem is to use the biological background knowledge already available to incorporate it into the learning process. In practice, prior knowledge from the existing genomic databases is used to bias the selection towards some features assumed more relevant [28]. Many knowledge-based approaches assume that if different predictive genes truly represent the same underlying biology, then perhaps is necessary to evaluate genes as members of gene pathways or biological processes, and use biological information to somehow guide the selection of

predictive genes. Ideally, one would like to have detailed gene pathways information, which can then be used to select genes with a potential causal link to the disease. This has largely not been possible due to the complexity of gene interactions. However, the problem of using biological information can be tackled in different ways, resorting to different types of external knowledge. Chuang *et al.* [29] use a mutual-information scoring approach to analyze known protein-protein interaction (PPI) networks, infer gene pathways, and find subnetworks predictive of breast cancer metastasis. Another approach, proposed in Svensson *et al.* [30], analyzes expression data from ovarian cancers based on gene sets from the Gene Ontology (GO) [31]; to represent the set's expression the proposed method uses a statistic that is essentially a majority-vote of the over- and under-expressed genes defined using correlation. Machine-learning strategies were proposed in [32], which, using the semantic structure of GO database, considers a set classifiers representing Gene Ontology terms instead of one overall classifier, and in [33], which uses pathway-specific regularization parameters (in Chapter 4 these last two methods will be discussed in detail).

In other approaches, biological information is not used to group variables, but rather to help the learning process to select features which are coherent in a biological sense. Kernel methods are among the most interesting (and recently devised) algorithms [34], *e.g.* Support Vector Machines [35]. In this kind of methods, different kernel functions can be plugged into any linear learning algorithm, *e.g.* the perceptron [36]. The kernel function is a formulation which efficiently represents a similarity function between two input objects with the property that this corresponds to a scalar product in an opportune (typically very high-dimensional) feature space. The most important property of these methods is that they allow to deal with structured objects very easily since one can change the kernel used (and then the discriminant function) but still keeping the same learning algorithm infrastructure. Examples of this type of approaches are presented in [37, 38]. In order to incorporate this information into methods for biomarker discovery, it is possible to use the matrix of pairwise similarities or dissimilarities of gene annotations to score the connections between genes. Several methods have been defined from the information theory to score the similarity between GO annotations [39]. Many tools have been developed to automatically use and mine this annotation, for an overview see [40]. In chapter 5 a detailed overview of these integration methods and similarity measures used on different types of biological knowledge is presented.

It has been shown that the use of prior knowledge on relevant features induces

a large gain in stability with improved classification performance [28, 41] and it seems a promising approach as the current genomic databases are significantly growing up providing new available a priori knowledge.

1.3 Mining the information in genomic databases

Although prior knowledge is helpful in improving the stability of feature selection, the use of such information deserves certain limitations caused by the knowledge domain represented by genomic databases. In the post-genomic era, public databases are the main place where one can deposit and/or screen the data. In the last few years, both the amount of electronically stored biological data and the number of biological data repositories have grown up significantly. Nucleic Acids Research magazine annually publishes the current database compendium in a special database issue. In the 2012 year issue, 92 new and 100 updated data resources are described. Database collection counts now 1380 different resources [42].

The only problem is to know how to benefit from this rich source of information. Thus, the accurate analysis of biological data and repositories turns out to be useful to obtain a systematic view of biological database structures, tools and contents and, eventually, to facilitate the access and recovery of such data. The availability of so many data repositories is an important resource but, on the other hand, opens new questions, as to effectively access and retrieve available data. Indeed, the heterogeneity of biological data and sources can often cause trouble to the user trying specific demands; moreover, not only such data sources are so numerous, but they use different kinds of representations, methods and various features and formats.

The fields of how to store and mine this knowledge in order to systematically incorporate it into clinical prediction algorithms are just beginning to develop. Different types of biological background knowledge exist depending on the level of the system that is described. First, a controlled vocabulary is necessary to annotate and to be able to systematically use this knowledge. The GO consortium represents an initiative to define a structured vocabulary to annotate and mine gene functions using a relational database where biological functions are hierarchically linked together. Some databases focus only on the biological interactions that characterize the processes within a living cell and summarize these interactions in usually manually curated pathway models, *e.g.* KEGG [43] or ConsensusPathDB [44]. Other databases focus only on certain types of molecular inter-

actions or on data obtained by certain genomics techniques such as transcription factor binding based on chip-chip data, *e.g.* TRANSFAC [45] or JASPAR [46], or protein-protein interactions based on co-immuno-precipitation or yeast two-hybrid screening, *e.g.* Human Protein Reference Database (HPRD) [47]. These types of biological knowledge can be represented in graph structures, where the vertices represent the genes or molecules and the edges represent some types of molecular interactions or relationships. In Chapter 3, a detailed description of Gene Ontology and PPI databases, used in the integration methods presented in this thesis, is provided.

The frequent addition of new genomes into public sequence databases allows an accumulation of information that is astounding in both its scale and breadth. While these data hold enormous promise for biological and medical discovery, experimental characterization has been performed on only a small fraction of the available sequences. As a result, computational methods are required to predict the molecular functions of the millions of protein sequences that have not and cannot be characterized experimentally. For over a decade, the majority of sequences found in public databases has been annotated using computational methods alone, raising the issue of annotation accuracy and database quality [48, 49].

In a recent paper that modeled annotation error in the Gene Ontology database, it was estimated that up to 49% of computationally annotated sequences could be misannotated [50]. Considering the problem from a different perspective, models of error propagation have shown that with sufficient initial error in a database, error propagation can significantly degrade the quality of the annotations it contains [51, 52] and specific examples of error propagation have been noted. Although functional misannotation remains an open issue [53], an in depth analysis of the prevalence of annotation error in large public databases has yet to be performed. Chapter 6 will discuss in detail the problem related to heterogeneities and possible inconsistencies of annotation, focusing on Gene Ontology database.

1.4 Thesis Overview

This thesis explores the intrinsic complexity of biomarker discovery task, in particular focusing on the reproducibility and interpretability of biomarker lists. This work started from a first study on an assessment of the main classification and feature selection methods, where particular learning strategies have been tested and evaluated in terms of reproducibility of results and prediction accu-

racy. Results highlighted that the bootstrap approach, which provides an ensemble output from different classifiers and different selected features sets, is able to significantly improve the reproducibility of the results. However, this approach can be applied regardless of the possible relationships among features, which can strongly affect both stability of results and lead to molecular signatures concretely useful for clinical studies. Indeed, a desirable property of a biomarker list is its interpretability in a biological sense, by selecting the most discriminating features which can be related to a specific biological process or pathway compromised by the disease.

Addressing this point, the core of the thesis deals with the integration of different types of prior knowledge in the learning process: the basic idea of this strategy is to take into account the complex gene relationships, instead of considering genes as independent features. In this work, two different kinds of approaches to integrate biological information are presented. Focusing on the biological information collected in the Gene Ontology database (GO), the first approach efficiently exploits the specific structure of the database defining meta-features which represent biological processes or molecular functions in order to provide a more interpretable list of those mechanisms altered by the disease. In the second proposed approach, different types of biological information like functional annotations, protein-protein interactions and expression correlation among genes were evaluated in the context of classification analysis and feature ranking by codifying each type of information into similarity matrices between features, in order to transform the feature space such that more similar two features are, the more closely they are mapped. In this way, an assessment of the effect of different types of biological knowledge was performed, in terms of both predictive accuracy and feature ranking stability.

In the integration of prior knowledge from genomic databases, the reliability of the biological information is important, because it strongly affects the ranking of the biomarker lists. Thus, in parallel with the investigation of the issue of biomarker list stability, the problem of mining the biological knowledge and handling the intrinsic heterogeneity of available annotations was analyzed. A global analysis of heterogeneity in the GO, which is the most popular functional annotation database used in genomic studies, was performed in order to study the presence of heterogeneities among the GO annotations and the impact of these heterogeneities on the extraction of biological information from this database. Moreover, a useful approach able to mining GO information accounting for these aspects was developed, in order to consider the quality and the origin of annota-

tions when biological annotations are investigated.
The two works presented in the appendices are slightly different with respect to the aim of this thesis, but useful to appreciate the advantages of the use of knowledge-based approaches into the analysis of high-throughput data. In particular, they describe the integration of Gene Ontology information in un-supervised classification methods in order to improve the functional characterization of temporal patterns of genes selected as differentially expressed.
In the following, a detailed description of the organization of the thesis is displayed.

Chapter 2 - Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment

In this chapter, an assessment of the consistency of candidate biomarkers provided by a number of different methods was performed, both on simulated (through an in silico regulation network model) and real clinical datasets. The effect of heterogeneity characterizing complex diseases was extensively simulated, reproducing both intrinsic variability of the population and the alteration of regulatory mechanisms. Population variability was simulated by modeling evolution of a pool of subjects; then, a subset of them underwent alterations in regulatory mechanisms so to mimic the disease state. Different methods for binary classification and feature weighting and ranking were applied to simulated datasets of different sample size to assess average classification performance and list stability: the Spectral Regression Discriminant Analysis algorithm, the classical Support Vector Machines and its variants. In all experiments, external bootstrap with separate training and test phases were employed to avoid overfitting effects such as selection bias. Results were also compared with those obtained by using SAM, a widely applied variant of univariate statistical t-test [54].
The simulated data highlighted advantages and drawbacks of different methods across multiple studies and varying number of samples and evaluated precision of feature selection on a benchmark with known biomarkers. Although comparable classification accuracy was reached by different methods, the use of bootstrap approach is helpful in finding features with a higher degree of precision and stability, preserving high classification accuracy. Application to real data confirmed these results. The material of this chapter partially appears in:

- DI CAMILLO B, SANAVIA T, MARTINI M, JURMAN G, SAMBO F, BARLA A, SQUILLARIO M, FURLANELLO C, TOFFOLO G, COBELLI C Effect of

size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment. PLoSOne (Accepted)

- DI CAMILLO B, MARTINI M, SANAVIA T, JURMAN G, SAMBO F, BARLA A, SQUILLARIO M, FURLANELLO C, TOFFOLO G, COBELLI C (2010) Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment. In: 9^{th} European Conference on Computational Biology (ECCB), 26-29 September 2010, Ghent, Belgium.

- DI CAMILLO B, MARTINI M, SANAVIA T, COBELLI C, TOFFOLO G (2010) In silico assessment of effect of size and heterogeneity of samples on biomarker discovery. In: Secondo Congresso Nazionale di Bioingegneria (GNB). 8-10 July 2010, Turin, Italy.

Chapter 3 - Biological knowledge in genomic databases

The accumulation of data produced by genome-scale research requires explicitly defined vocabularies to describe the biological attributes of genes and their products in order to allow integration, retrieval and computation of data. Thus biological databases, which represent these vocabularies, are an important tool in bioinformatics to better understand a host of biological phenomena concerning the structure of biomolecules and their interactions, their molecular functions and the biological processes where they interact. This knowledge is collected from scientific experiments, published literature, high-throughput experiment technology, computational analyses and the information contained in biological databases includes gene functions, structures, localizations (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures. In some cases, biological information is organized according to relational concepts, given well-defined semantics in order to handle knowledge computationally in a manner comparable to numeric data. In particular, ontologies are used as a useful tool to capture the main concepts in a specific knowledge domain and are able to structure them in a systematic way by defining a set of entities with specific attributes and relationships among them. The most successful example is the Gene Ontology (GO), which describes biological processes, molecular functions and cellular components of gene products in both a computer- and human-readable manner.
Another useful information derives from the pathway databases, since gene products do not act independently, but in a network of complex molecular interactions. For example, signals from the exterior of a cell are mediated to the inside of a cell

by protein-protein interactions of the signaling molecules. Methods for identifying interacting biological components, such as proteins and metabolites, have defined hundreds of thousands of interactions. These interactions are collected together in specialized biological databases that allow the interactions to be assembled and studied further. In this chapter, a detailed description of both Gene Ontology, pathways and protein-protein knowledge is provided.

Chapter 4 - Gene Ontology based classification: improving prediction and biological interpretability

The standard biomarker discovery applications assume that all features are equally relevant before the selection procedure. In practice, since only a subset of features are relevant for the biological case analyzed, some prior knowledge may be available to bias the selection towards these features. Many genes are known to have the same function or involved in the same biological process as some known/putative disease-related genes, and the genes in the same functional group are more likely to work together. Functional analysis was introduced to address a better biological characterization of results. In biomarker discovery, some important aspects significantly affect the biological characterization of biomarker lists and are still now an open issue and are only partially dealt by the currently available classification and feature selection methodologies: 1) the biological information, *i.e.* annotations from Gene Ontology, often considered only a posteriori, without affecting the gene signature extraction; 2) the correlation among the features, reflecting the combined effect of multiple features on disease; 3) the organization of results in a structured, easy-to-read way in order to achieve a better interpretation of biological processes altered by the disease. Considering functional annotations of GO database, in this chapter a new method is presented, which is able to integrate classification/feature selection with functional annotations in order to increase biological interpretability of gene signatures by defining subsets of genes both correlated and annotated to groups of GO terms with similar meaning. The improvements on both classification performance and reproducibility of selected GO terms across different datasets suggest narrowing the search of biomarkers among a limited set of genes characterized by similar functional roles or interacting in the same biological process. The material of this chapter partially appears in:

- SANAVIA T, CREPALDI A, BARLA A, DI CAMILLO B (2011) Gene Ontology based classification improves prediction and gene signature interpretability. Net-

work Tools and Applications in Biology (NETTAB) Workshop, 12-14 October 2011, Pavia, Italy.

- SANAVIA T, BARLA A, DI CAMILLO B, MOSCI S, TOFFOLO G (2009) Function-based analysis of microarray data via l1-l2 regularization. In: 17^{th} Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 8^{th} European Conference on Computational Biology (ECCB), 27 June - 2 July 2009 Stockholm, Sweden.

Chapter 5 - Gene Ontology based classification: prediction and biological information stability and interpretability

A possible approach to improve list stability is to integrate biological information from genomic databases in the learning process; however, a comprehensive assessment based on different types of biological information is still lacking in the literature. In this chapter, a comparison of the effects of using different biological information in the learning process like functional annotations, protein-protein interactions and expression correlation among genes was performed. Biological knowledge was codified by means of gene similarity matrices and expression data linearly transformed in such a way that the more similar two features are, the more closely they are mapped. Two semantic similarity matrices, based on Biological Process and Molecular Function Gene Ontology annotation, and geodesic distance applied on protein-protein interaction networks, are the best performers in improving list stability maintaining almost equal prediction accuracy. The performed analysis supports the idea that when some features are strongly correlated to each other, for example because they are close in the protein-protein interaction network, then they might have similar importance and are equally relevant for the task at hand. The performance of different sources of prior knowledge was evaluated using three real datasets from different studies exploring the same clinical classification task. The assessment of the results obtained for different similarity matrices is based on the trade-off between predictive accuracy and feature ranking stability. Obtained results can be a starting point for additional experiments on combining similarity matrices in order to obtain even more stable lists of biomarkers. The material of this chapter partially appears in:

- SANAVIA T, AIOLLI F, DA SAN MARTINO G, BISOGNIN A, DI CAMILLO B Improving biomarker list stability by integration of biological knowledge in the learning process. BMC Bioinformatics, Volume 13, Supplement 3, 2012.

- SANAVIA T, AIOLLI F, DA SAN MARTINO G, BISOGNIN A, DI CAMILLO B (2011) Improving biomarker list stability by integration of biological knowledge in the learning process. In: 19^{th} Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 10^{th} European Conference on Computational Biology (ECCB), 15-19 July 2011, Wien, Austria.

- SANAVIA T, AIOLLI F, DA SAN MARTINO G, BISOGNIN A, DI CAMILLO B (2011) Stable Feature Selection for Biomarker Discovery: Use of Biological Information. In: BITS Annual Meeting 2011, 20-22 June 2011, Pisa, Italy.

Chapter 6 - Gene Ontology based classification: prediction and biological information stability and interpretability

In the integration of prior knowledge from genomic databases the reliability of the biological information is an important issue, because it strongly affects the ranking of the biomarker lists. Referring to Gene Ontology (GO), which is the most widely used annotation database to transfer biological knowledge on gene product, in this chapter a comprehensive study of the global level of heterogeneity in GO annotations was carried out, in order to quantify to what extent inconsistencies are present in the GO database. In particular, the work focused on protein annotations belonging to Biological Process (BP) and Molecular Function (MF), the most frequently used GO categories. The global analysis of inconsistencies in the GO database revealed that the percentage of groups of proteins with high sequence similarity but non-homogeneous annotations is around 10% for Biological Process and 8% for Molecular Function. These obtained results are indicative of the presence of heterogeneities and confirm the need of considering the quality and the origin of GO annotations when inferring possible biological functions. To this purpose, it was developed a method able to assess annotations related to pools of proteins sharing similar biological functions, in order to organize results into a functional map as a useful guidance to easily interpret the reliability of GO annotations. The material of this chapter partially appears in:

- FACCHINETTI A, SANAVIA T, DI CAMILLO B, LAVEZZO E, FONTANA P, TOPPO S (2011) A Method to Reveal and Handle Heterogeneities and Inconsistencies in Gene Ontology Annotation. In: BITS Annual Meeting 2011, 20-22 June 2011, Pisa, Italy.

- SANAVIA T, FACCHINETTI A, DI CAMILLO B, TOFFOLO G, LAVEZZO E, TOPPO S, FONTANA P (2010) Revealing heterogeneities and inconsistencies in

protein functional annotations. In: 9^{th} European Conference on Computational Biology (ECCB). 26-29 September 2010, Ghent, Belgium.

Appendices

The two works presented in the appendices describe two different approaches to exploit and integrate functional information in gene expression analysis, focusing on the un-supervised classification, *i.e.* clustering methods. In the first study, a new workflow to analyze dynamic gene expression data was defined and the pipeline was applied to study the effect of insulin on human endothelial cells, providing new insights on how insulin affects different biological functions, characterized by well-defined temporal patterns. The second study proposed a method to integrate the three main analyses usually performed on dynamic gene expression data (gene selection, clustering and functional interpretation) addressing different drawbacks affecting these analyses. The method, applied on skeletal muscle cells treated with insulin, allowed identifying characteristic dynamic responses to insulin stimulus, common to a number of genes and associated to the same functional group. The material of the appendix partially appears in:

- DI CAMILLO B, SANAVIA T, IORI E, BRONTE E, RONCAGLIA E, MARAN A, AVOGARO A, TOFFOLO G, COBELLI C (2010). The Transcriptional Response in Human Umbilical Vein Endothelial Cells Exposed to Insulin: a Dynamic Gene Expression Approach. PLoSONE 5(12):e14390.

- SANAVIA T, DI CAMILLO B, IORI E, MARAN A, BRONTE E, AVOGARO A, TOFFOLO G, COBELLI C (2008). Function-based discovery of characteristic temporal expression profiles in endothelial cells stimulated with insulin. In: 11^{th} International Meeting of Microarray and Gene Expression Data Society (MGED), 1-4 September 2008 Riva del Garda (TN), Italy.

- DI CAMILLO B, IRVING BA, SCHIMKE J, SANAVIA T, TOFFOLO G, COBELLI C, NAIR KS Function-based discovery of significant transcriptional temporal patterns in insulin stimulated muscle cells. (Under second revision)

Chapter 2

Effect of size and heterogeneity of samples on biomarker discovery

2.1 Background

Gene expression data from microarrays have been extensively used both to predict pre-clinical and clinical endpoints and to identify the most discriminating features representing biomarkers indicating pathogenic processes, or potential targets of clinical and pharmaceutical applications. In order to address these two issues, different supervised classification and feature selection methods have been developed: the former, applied to patients based on gene expression measurements, have proved to be useful for answering to several diagnostic/prognostic questions and a variety of predictive models have been suggested; the latter provide different approaches to avoid overfitting and improve model performance, have found a direct application on biomarker discovery task to gain a deeper insight into the underlying processes generating the data.
However, there are several questions related to the application of both classification and feature selection methods on high-throughput data which are still matters of several scientific discussions since, as pointed out in the first chapter, one of the properties of microarray data is that there are many predictors (genes) with a small sample size. In the context of classification analysis, high-dimensional data strongly affect the performance in terms of accuracy on assigning the correct class or on predicting a clinical parameter for new patients, because predictive models have to cope with large amounts of irrelevant features. In the last decade, several different methods like discriminant analysis, random forests and support vector machines among others, have been used on gene expression data, especially

in cancer studies [55, 56, 57], where early diagnosis and an accurate classification are crucial for treatment or therapy. Although the performance of these classifiers can be very high in the studied dataset, application of these predictive models in other datasets is often limited and data reproduction is not straightforward [58]. On the other hand, since classic feature selection methods aim at selecting the minimum subset of features to construct a classifier with the best predictive accuracy, the existing methods often ignore the stability issue, because they do not necessarily identify the same candidate feature subsets if the biomarker discovery procedure is repeated [20]. Even for the same data, one may find many different subsets of features (either from the same feature selection method or from different feature selection methods) that can achieve the same or similar predictive accuracy [21, 22, 23]. Boutros *et al.* [59] showed that the use of different statistical procedures could identify multiple highly prognostic signatures from one dataset [60]. An extensive analysis of the effect of different statistics on ranked gene lists showed large variability [18]. In particular, Ein-Dor *et al.* [13] suggested that many more samples than currently available would be required to reach a good level of signature stability. The stability issue in feature selection has received much attention recently and several works have recently pointed out that high reproducibility of biomarkers lists is equally important as high classification accuracy [61, 62, 19].

A slightly different issue, although related with list stability, is the precision of biomarker identification, *i.e.* the ability to select true biomarkers, defined as features biologically related to the physiological or clinical condition under study as cause or effect of it. It is possible that there exist multiple sets of potential true markers in real data. When there are many highly correlated features, different ones may be selected under different settings [20]. On the other hand, even there are no redundant features, the existence of multiple non-correlated sets of real markers is also possible [63].

Different studies were proposed to assess the current status of biomarker discovery methods. In a recent study [55], classification performance of different methods was compared across different microarray studies in terms of ability to select biomarkers discriminating between two conditions. Addressing also the stability issue, the recent work presented in [64] compares 32 feature selection methods on four public gene expression datasets for breast cancer prognosis, in terms of both predictive performance and stability, demonstrating that complex multivariate methods, able to identify multivariate nonlinear combinations of features and thus the interplay between regulated genes, generally do not outperform simple

univariate feature selection methods, which instead consider genes as independent features.

2.2 An univariate approach: Statistical analysis of Microarrays (SAM)

The SAM test [54] is a widely used univariate statistical test for the identification of differentially expressed genes from microarray data. Gene expression data on p genes for n mRNA samples may be described by an $n \times p$ matrix X=(x_{ij}), where x_{ij} denotes the expression level of gene (variable) j in mRNA sample (observation) i. SAM identifies statistically significant genes by carrying out a variant of t-test using a statistic $d(i)$ for each gene i which measures the relative difference between samples related to the gene expression level $x(i)$:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0} \tag{2.1}$$

where $\bar{x}_I(i)$ and $\bar{x}_U(i)$ are defined as the average levels of expression for gene i in states I and U, respectively. The "gene-specific scatter" $s(i)$ is the standard deviation of repeated expression measurements:

$$s(i) = \sqrt{a \left\{ \sum_{m=1}^{N_1} [x_m(i) - \bar{x}_I(i)]^2 + \sum_{n=1}^{N_2} [x_n(i) - \bar{x}_U(i)]^2 \right\}} \tag{2.2}$$

where $a = (1/N_1 + 1/N_2)/(N_1 + N_2 - 2)$, with N_1 and N_2 indicating the number of measurements in the states I and U. The $d(i)$ or relative difference between samples for a given gene includes s_0, which corrects for high variations in samples with relatively low intensities, which is typical for microarray data.
SAM uses a resampling procedure to derive the null hypothesis distribution and the false discovery rate (FDR), *i.e.* the expected number of false positives within a group of positives, to account for multiple testing [5]. In this study, a FDR=5% was used to select features after a ranking based on their p-value. The use of permutation-based analysis accounts for correlations in genes and avoids parametric assumptions about the distribution of individual genes. This is an advantage over other techniques which assume equal variance and/or independence of genes.

2.3 Classification and feature selection methods

Before describing the classification/feature selection methods used in this study, in the following, some useful terms and notations are first introduced. When the mRNA sample belong to known classes, the data for each observation consist of a gene expression profile $\mathbf{x}_i$=(x_{i1},...,x_{ip}) and a class label y_i. For K classes, the class labels y_i are defined to be integers ranging from 1 to K. In the following, binary classification problems are considered (K=2). A *classifier* for K classes partitions the space of the gene expression profiles into K disjoint subsets, A_1,...,A_K, such that for a sample with expression profile $\mathbf{x}$=(x_1,...,x_p)$\in A_k$ the predicted class is k.
Classifiers are built from past experience, *i.e.* from observations which are known to belong to certain classes. Such observations comprise the *training set* (TR) $\mathcal{L}$=$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n_L}, y_{n_L})\}$. Classifiers may then be applied to a *test set* (TS) $\mathcal{T}$=$\{\mathbf{x}_1, ..., \mathbf{x}_{n_T}\}$, to predict for each observation $\mathbf{x}_i$ in the test set its class y_i. In the case that y_i is known, the predicted and true classes may be compared to estimate the error rate of the classifier. Considering the entire gene expression matrix X, $\mathbf{y}$=(y_1,...,x_n) indicates the vector of observations.

2.3.1 Discriminant Analysis

Linear Discriminat Analysis (LDA) method was originally proposed by Fisher [65] as a means for finding the optimal linear combination of variables able to minimize the within-class distance and to maximize the between-class distance simultaneously, in order to achieve the maximum class discrimination.
Considering an $n \times p$ gene expression data matrix X, the method finds linear combinations $\mathbf{xa}$ of the gene expression levels $\mathbf{x}$=(x_1,...,x_p) with large ratios of between-groups to within-groups of sum of squares:

$$\hat{\mathbf{a}} = argmax_{\mathbf{a}} \frac{\mathbf{a}^T S_b \mathbf{a}}{\mathbf{a}^T S_w \mathbf{a}} \tag{2.3}$$

where S_b and S_w are respectively the $p \times p$ between group and within group (pooled) covariance matrices of the gene expression values. When l projective functions A=[$\mathbf{a}_1$,...,$\mathbf{a}_l$] are needed, the objective function of LDA can be written as:

$$\hat{A} = argmax_A \frac{tr(A^T S_b A)}{tr(A^T S_w A)} \tag{2.4}$$

where $tr(.)$ denotes matrix trace. Applying an eigen-decomposition on these matrices derived from the learning set, the optimal transformation can be obtained.

The optimization problem in Equation 2.4 is equivalent to finding the l eigenvectors of the following generalized eigen-problem associated with maximum eigenvalues:

$$S_b\mathbf{a} = \lambda S_w\mathbf{a} \tag{2.5}$$

A common problem of this method is the high computational cost from the eigendecomposition of dense matrices. To address this problem, Cai *et al.* [66] introduced Spectral Regression Discriminant Analysis (SRDA), which casts discriminant analysis into a regression framework by using spectral graph analysis: the regression framework improves the computational efficiency, whereas the spectral graph analysis is used for solving a set of regularized least squares problems avoiding the eigenvector computation. In particular, the method finds $\mathbf{a}$ which satisfies $X^T\mathbf{a}$=$\mathbf{y}$ by a regularized form of the ordinary least squares estimator:

$$\hat{a} = (XX^T + \alpha I)^{-1}X\mathbf{y} \tag{2.6}$$

where $\alpha(\geq 0)$ is the only regularization parameter needed to be tuned, which controls the smoothness of the estimator $\hat{a}$.

2.3.2 Support Vector Machines

If each vector in the gene expression matrix is considered as a point in an n-dimensional space, a simple way to build a binary classifier is to construct a hyperplane separating class members from non-class members in this space. Unfortunately, most real-world problems involve non-separable data for which does not exist a hyperplane that successfully separates the class members from non-class members in the training set. One solution to the inseparability problem is to map the data into a higher-dimensional space and define a separating hyperplane there. This higher-dimensional space is called the *feature space*, as opposed to the *input space* occupied by the training examples. With an appropriately chosen feature space of sufficient dimensionality, any consistent training set can be made separable. However, representing the feature vectors corresponding to the training set can be extremely expensive in terms of memory and time. Furthermore, the artificial separation of the data using this approach leads to obtain trivial solutions that overfit the data. To address these problems, Support Vector Machines (SVM) were proposed [35].

Given training vectors $\mathbf{x}_i$=(x_{i1},...,x_{ip}) and its corresponding class $y_i \in \{-1; 1\}$, the support vector technique tries to find the separating hyperplane with the largest margin between two classes, measured along a line perpendicular to the hyperplane. For example, in Figure 2.1, two classes could be fully separated by a dashed

line $\mathbf{w}^T\mathbf{x}_i+b=0$. The SVM approach computes a hyperplane that maximizes the margin separating the two classes of samples. The optimal hyperplane is called *decision boundary*. That means the method finds a line with parameters $\mathbf{w}$ and b such that the distance between $\mathbf{w}^T\mathbf{x}_i+b=\pm 1$ is maximized. By rescaling the

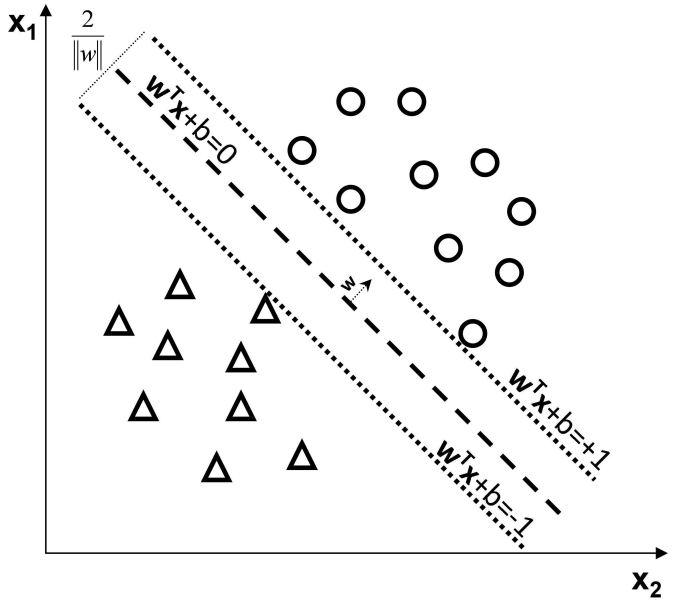

Figure 2.1: **Example of separating hyperplane.**

parameters $\mathbf{w}$ and b, the margin d can be written as $d=2/\|\mathbf{w}\|$. The learning task in SVM can be formalized as the following constrained optimization problem:

$$min_{\mathbf{w}} = \frac{\|\mathbf{w}\|^2}{2} \tag{2.7}$$

subject to $y_i(\mathbf{w}^T\mathbf{x}_i+b)\geq 1$.

When the classes are not linearly separable, a variant of SVM, called soft-margin SVM, is used. This SVM variant penalizes misclassification errors and employs a parameter (the soft-margin constant C) to control the cost of misclassification. Thus, training a linear SVM classifier amounts to solving the following constrained optimization problem:

$$min_{\mathbf{w},b,\xi}\frac{\|\mathbf{w}\|^2}{2} + C\sum_{i=1}^{n}\xi_i \tag{2.8}$$

with one constraint for each training sample $\mathbf{x}_i$: $y_i\mathbf{w}^T\mathbf{x}_i+b\geq 1\text{-}\xi_i$, where ξ_i are defined *slack variables*. These constraints allow that training data may not be on the correct side of the separating hyperplane $\mathbf{w}^T\mathbf{x}_i+b=0$ while the training error

$\sum_{i=1}^{n} \xi_i$ is minimized in the objective function. Hence, if the penalty parameter C is large enough and the data is linear separable, all ξ_i will be zero [67].
Usually, the dual form of the optimization problem is solved:

$$min_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j \tag{2.9}$$

subject to $\alpha_i \geq 0$,$\sum_{i=1}^{n} \alpha_i y_i$=0, where α_i is a real number. The resulting decision function of a new sample $\mathbf{z}$ is $f(\mathbf{z})$= $\mathbf{w}^T\mathbf{x}_i$+b with $\mathbf{w}$=$\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ and b=<y_i-$\mathbf{w}$·$\mathbf{x}_i$ >. Usually many α_i are zero. The training samples $\mathbf{x}_i$ with non-zero α_i are called *support vectors*. The weight vector $\mathbf{w}$ is a linear combination of support vectors. The bias value b is an average over support vectors. The class label of $\mathbf{z}$ is obtained by considering the sign of $f(\mathbf{z})$.
The decision function for classifying points with respect to the hyperplane only involves dot products between points in the feature space. Because the algorithm that finds a separating hyperplane in the feature space can be stated entirely in terms of vectors in the input space and dot products in the feature space, a support vector machine can locate the hyperplane without ever representing the space explicitly, simply by defining a function, called the $kernelfunction$, that plays the role of the dot product in the feature space. This technique avoids the computational burden of explicitly representing the feature vectors. Moreover, since the simple inner product does not always measure the similarity effectively for all applications, for some applications a non-linear decision boundary is more effective for classification. The basic SVM method can then be extended by transforming samples to a higher dimensional space via a mapping function ϕ. By doing this, a linear decision boundary can be found in the transformed space if a proper function ϕ is used. The kernel function overcomes the limitations related to the computation in the transformed space, which can be expensive because of its high dimensionality. The kernel function can defined as $K(\mathbf{x}_i,\mathbf{x}_j)$=$\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ denote the i-th and j-th sample, respectively. Different kernels can be used: linear ($K(\mathbf{x}_i,\mathbf{x}_j)$=$\mathbf{x}_i^T\mathbf{x}_j$), polynomial with degree γ ($K(\mathbf{x}_i,\mathbf{x}_j)$=$(1+\mathbf{x}_i^T\mathbf{x}_j)^\gamma$), Gaussian ($K(\mathbf{x}_i,\mathbf{x}_j)$=$exp^{-\|\mathbf{x}_i-\mathbf{x}_j\|/2\sigma^2}$); in this last case, each data point is mapped to a gaussian function with bandwidth σ, thus the hyperplane is a combination of gaussian functions of the support vectors.
In this study, SVM method was implemented using both linear (LSVM) and Gaussian kernel (GSVM). The tuning phase required the identification of the optimal value of the regularization parameter C (the trade-off between empirical error and smoothness of the solution) and, for the Gaussian kernel, of the bandwidth σ.

2.3.3 Ranking and selection

Besides the predictive models, also an efficient approach to rank and extract the most discriminant features is important for biomarker discovery task. For feature ranking, the coefficients $\mathbf{a}_i$ of SRDA and the weights $\mathbf{w}_i$ of a SVM classifier provide information about feature relevance. A variant of feature ranking for linear SVM used in this study is the Iterative-Relief [68], which assigns a score to features based on how well the features separate training samples from their nearest neighbours from the same and from the opposite class. The method solves a convex optimization problem with a margin-based objective function in a nearest-neighbor based strategy. The algorithm constructs iteratively a weight vector, which is initially equal to zero; at each iteration, it selects one sample, adds to the weight the difference between that sample and its nearest sample from the opposite class (called *nearest miss*), and subtracts the difference between that sample and its nearest neighbour from the same class (called *nearest hit*). The iterative process terminates when all the training samples have been considered. The ranking provided by Iterative-Relief can be used by an independent classifier: in this study, it was used together with linear SVM (IRSVM).
Another task is the selection of the most discriminant feature. The Recursive Feature Elimination (RFE) is a well-known feature selection method, firstly introduced in [69]. Considering its application on the SVM-based prediction models, the method defines the importance of a feature for a SVM in terms of its contribution to the cost function $J(\alpha)$, defined in Equation 2.9. At each step of the RFE procedure, a SVM is trained on the given data set, J is computed and the feature less contributing to J is discarded. In the case of linear SVM, the variation due to the elimination of the i^{th} feature is $\delta J(i)=\mathbf{w}_i^2$. The heavy computational cost of RFE is a function of the number of variables, because a SVM must be trained each time a variable is removed. However, at the first loops of the RFE algorithm, many weights are generally similar and concentrated nearby zero [70]. In the standard RFE algorithm just one of the many features corresponding to a minimum weight would be eliminated, while it would be convenient to remove all of them at once. Another possible choice is to remove $\left\lfloor \sqrt{|R|} \right\rfloor$ features at each step, where R is the set of the remaining features, thus obtaining the SQRT-RFE procedure. Furlanello *et al.* [70] developed an *adhoc* strategy for an elimination process based on the structure of the weight distribution, using an entropy function H, namely the Entropy-based Recursive Feature Elimination (E-RFE). To compute the entropy, the range of the weights are split into n_{int} intervals, with

$n_{int} = \left\lfloor \sqrt{|R|} \right\rfloor$; for each interval, the relative frequency is defined as:

$$p_i = \frac{|\delta J(i)|}{|R|} \tag{2.10}$$

Entropy is then defined as the following function:

$$H = -\sum_{i=1}^{n_{int}} p_i log_2 p_i \tag{2.11}$$

According to the entropy measure of the distribution of the weights generated by the feature weighting schema, E-RFE adaptively discards a subset of the least informative features, speeding-up the ranking procedure without performance degradation.
In this study, for all the considered classification methods, E-RFE procedure was used as the ranking schema and the optimal number of features was chosen in correspondence to the minimum classification error estimate.

2.4 Cross validation and bootstrap approach

For assessing the performance of a classifier, a validation phase, which is independent with respect to the generation of the predictive model from the training set of samples, is perfomed using the samples belonging to the test set. The correctly classified proportion of samples in a test set is an estimate of accuracy assessing the classifier performance. If there is an abundance of data, this estimate is unbiased and variance tends to zero as the number of test samples goes to infinity. The most direct approach is a split sample validation method, which randomly splits samples into a training set and a test set.
The k-fold cross validation method partitions the samples into k non-overlapping subsets of as close to equal size as possible, assigning k-1 subsets into a training set and the remaining subset into a test set, in order to develop the prediction classifier using the training set and to estimate its accuracy using the test set. After iterating the procedure k times until all subsets have had a chance to be a test set, every sample has an associated cross-validated predicted class membership in addition to a true class membership, and the proportion of correctly classified samples among all available samples provides the estimate of accuracy. In the case of k=n, where each observation (microarray sample) in the training set is deleted in turn before it is allocated by the classifier built from the remaining n-1 observations, the so-called $leave - one - out$ cross validation is performed.

Regardless of how the performance of the classifier is assessed during the feature-selection process, it is common to assess the performance of the predictive model for a selected subset of genes by its cross-validated (CV) error. But, if it is calculated within the feature-selection process, there is a selection bias in it when it is used as an estimate of the prediction error [71]. Thus, an external split sample should be undertaken subsequent to the feature-selection process to correct for this selection bias. In this study, an external bootstrap approach was adopted. Given a standard training set TR of size m, bootstrap generates B new training sets D_i, each of size $m'=m$, by sampling examples from D uniformly and with replacement. The B models are fitted using the above B bootstrap samples and combined by averaging the output (for further details, see [70]).
The four methods considered in this analysis, LSVM, GSVM, IRSVM and SRDA, were used both in single cross-validation and in a MonteCarlo bootstrap resampling schema with $B=100$ external training/test splits with 3-fold cross-validation as internal sampling schema (methods named as LSVM_B, GSVM_B, IRSVM_B and SRDA_B in the following).

2.5 Simulation of gene expression data

In order to assess classification methods' performance across multiple studies with varying number of samples and to evaluate precision of feature selection on a benchmark with known biomarkers,microarray data were simulated starting from a gene network simulator described in [72]. Each simulated subject was modeled by a regulatory network of $N=10000$ genes, based on the gene network simulator using default parameter settings. Network topology was randomly generated with scale-free distribution of node degree and clustering coefficient independent of the number of nodes. Metabolic [73] and possibly transcriptional [74] networks exhibit this topological property. The topology is characterized by the connectivity matrix W, with weights w_{ij} different from zero if gene-product j directly affects the expression of gene i. Weights w_{ij} of the connectivity matrix W can be interpreted as the affinity of the genome specific sequences for a transcription factor or an enhancer j, regulating expression of a gene i. Since weights w_{ij} can in principle be mapped to specific sequences in the genome, W can be interpreted as part of the **genotype**[1] of the subject. The sign and the magnitude of w_{ij} indicate the

[1]Genotype accounts for the genetic constitution of an individual, that is it refers to the coded, inheritable information represented by the set of genes in the DNA sequence (the genetic code), copied at the time of cell division or reproduction and are passed from one generation to the

sign and the strength of the regulation, respectively. A target value $T_i(W,t)$ was derived for gene i at time t as a function of the different action of its regulators ($T_i(W,t)$ represents the expression value to which gene i tends at time t as an effect of the expression level of its regulators). By explicitly representing interactions among the regulators of each gene, the simulated systems were characterized by a finite number of basins of attractors, where each attractor corresponds to a steady state or a periodic behavior. Since each network is characterized by a finite number of steady states, reachable from a specific set of initial conditions and/or external stimuli, each steady state can be interpreted as the **phenotype**[2] of an individual in a particular environmental condition.
Differential equations were used to model the dynamics of transcription and degradation as continuous variables and to describe transcription delay with different time constants for each gene. In particular, the rate of change of the expression level of gene i at time t was described as:

$$\frac{dx_i(t)}{dt} = \lambda_i \cdot [S_i(T_i(W,t), \alpha_i, \beta_i) - x_i(t)] \tag{2.12}$$

where λ_i is a time constant influencing both the rate of transcription and the degradation term and $S_i(T_i(W,t), \alpha_i, \beta_i)$ is a sigmoid activation function, modulating the target value $T_i(W,t)$ and depending on gene i specific parameters α_i and β_i:

$$S_i(T_i(W,t), \alpha_i, \beta_i) = \frac{1}{1 + exp[-\alpha_i \cdot (T_i(W,t) - \beta_i)]} \tag{2.13}$$

At the end, each subject is characterized by a specific genotype (the connectivity matrix W with weights w_{ij}) and a specific phenotype, thus a vector $\mathbf{x}=(x_1, ..., x_N)$ representing the steady states expression values of the N genes, obtained by solving the differential equation 2.12.

2.5.1 Simulation of population evolution and variability

Starting from this simulation schema, evolution of a population of M=1000 individuals was simulated using a procedure similar to the one described in [75]. Given specific initial conditions (*i.e.* environment condition which was fixed for the purpose of this work), the initial population at generation 1 consisted of

next (inheritable).

[2]Phenotype refers to the observed traits or anatomical features of an individual, such as structure, physiology and behavior. These physical attributes and behavioral characteristics determine the organism's ability to survive and reproduce in the environment. The genotype of an organism determines the phenotype of an organism to a large extent.

M individuals with identical connectivity matrix W and with N dimensional vectors of expression values obtained by considering the steady state reached by the system. Gene specific kinetic parameters α_i and β_i were sampled from Gaussian distributions with means μ_α, μ_β and standard deviations σ_α, σ_β. For each subject, μ_α and μ_β were set to 20 and 0.2, respectively, whereas σ_α and σ_β were sampled from a Gaussian distribution with means 0.5 and 0.02 and standard deviations equal to 0.075 and 0.0025, respectively. Parameters values (Equations 2.12 and 2.13) were empirically chosen so to generate in silico data with statistical distribution similar to those observed on the real datasets. To introduce genotype variability in the population, subsequent generations were produced by iteration of three steps: random pairing of individuals, mutation of a randomly chosen subset of subjects and selection of the surviving subjects. For computational reasons, these three steps were applied only to a sub-network of size N=900, indicated as W_{900} in the following, which was constrained to be not connected to any of the other 9100 nodes in the network. Each step is described in detail in what follows:

- *Pairing.* Offspring was created by randomly selecting two parents among the current population of M individuals and randomly combining rows of the connectivity matrix W_{900} from each parent with equal probability.

- *Mutation.* Mutation was simulated by changing each nonzero w_{ij} (which, by simulation, resulted equal to 1619 elements on a matrix of 900×900=810000 elements) with probability 0.025/1619. The new value of each mutated w_{ij} was sampled from a Gaussian distribution with mean and standard deviation equal to 0 and 1, respectively. Therefore, at each iteration, each subject mutated with probability 0.025.

- *Selection.* Assuming, in a naïve simplification of reality, that individuals behaved as haploid[3] organisms and that the initial phenotype was essential for survival, subjects with at least one mutated w_{ij} were allowed to survive only if their phenotype did not change with respect to the original population. In practice, the Euclidean distance between the expression profile of each mutated subject was calculated (the N dimensional vector of gene expression values at steady state) and the average expression profile

[3]Haploid is the term used when a cell has only one set of chromosomes. A normal eukaryote organism is composed of diploid cells, one set of chromosomes from each parent. Having haploid genetics means an organism has only one version of each gene in their DNA, one from the mother OR one from the father.

of subjects at generation 1; if Euclidean distance exceeded the value of 0.81 (corresponding to the percentile 99.5 of the observed distances) the subject was eliminated, otherwise he/she survived. At each generation, M individuals were generated, independently of the number of parents survived in the previous generation. Evolution proceeded for a time sufficient to have a final population of M subjects with the same phenotype but different genotype, *i.e.* 150 generations (Figure 2.2).

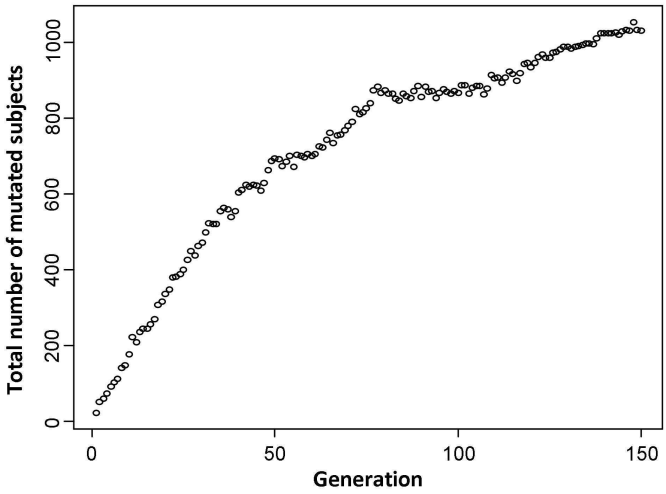

Figure 2.2: **Total number of subjects mutated with respect to the original population with the progress of generations.** Only survived subjects are represented for each generation.

Noise was added to expression data of the 10000 genes in the 1000 subjects as additive Gaussian noise with mean 0 and standard deviation sampled from the distribution of within-groups error variance in real datasets, as described in Di Camillo *et al.* [76]. In particular, the error variance associated to genes was approximated by a lognormal distribution with mean 0.22 and standard deviation 0.35.

2.5.2 Simulation of the pathological state

Once the base population was simulated, two groups, each of 500 subjects, were defined. For one of them a pathological condition is simulated by introducing some impaired regulation in the pathways of the non-healthy subjects. A possible strategy to do that is either to apply a gene silencing (*knock out*) or to reduce the expression level (*knock down*) of some genes. Thus, the pathological condition

was simulated by knocking out or down six target hubs, defined as those genes with the highest out-degree and expression value at steady state higher than 0.88, so that their knock out (down) achieved an effect. The knock out of gene j was simulated by setting to 0 its expression and all the elements of row j in matrix W. Consistently, the knock down of gene j was simulated by halving its value and all the elements of row j in matrix W. Diseased subjects had 4, 5 or 6 genes belonging to W_{900} that were knocked out or down. The proportion of subjects with 4, 5 or 6 genes affected was set equal to 1/3, 1/3, 1/3, respectively. For each gene, the proportion of subjects affected by knock-out and knock-down was set equal to 1/3 and 2/3, respectively.

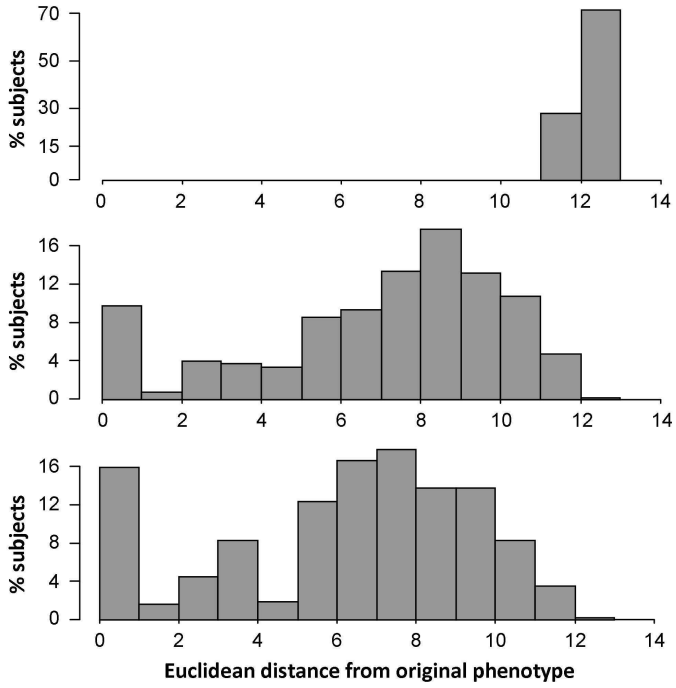

Figure 2.3: **Distance from the original phenotype of the evolved population.** Upper panel: after the knock-out of the six genes with the highest out-degree. Middle panel: after the knock out or knock down of all genes in proportion 1/3 and 2/3, respectively. Lower panel: after the knock out (down) of four, five or six genes.

Figure 2.3 displays group variability in terms of histogram of the Euclidean distance between the steady states of the original and the diseased population. The variability rises from both the intrinsic population variability, *i.e.* the different connectivity weights w_{ij} in W_{900}, and the heterogeneity of the disease. Figure 2.3, upper panel, shows the effect of the knock-out of the six hubs on subjects with

different connectivity weights w_{ij}, thus with the differences among subjects rising only from intrinsic population variability. Figure 2.3, middle panel, shows how these differences increase if hubs are knocked down rather than out in different subjects with frequency 2/3. Figure 2.3, lower panel, shows the diseased group variability, obtained when heterogeneity of the disease is also simulated: diseased subjects have 4, 5 or 6 affected genes (knocked out or down) in proportion 1/3, 1/3, 1/3, respectively. For each gene, the proportion of subjects affected by knock-out and knock-down was set equal to 1/3 and 2/3, respectively.
Comparison between simulated and Affymetrix data (GSE2990, see below) showed that the datasets have very similar distribution (Wilcoxon test p-value equal to 0.9).

2.5.3 Biomarkers definition and generation of simulated datasets

One of the main aspects of using simulated data is to evaluate precision of feature selection on a benchmark with known biomarkers. The expected biomarkers are defined as those genes directly or indirectly regulated by at least one of the six hubs, having expression modified by the knock out (down). Before simulating the pathological state, for each gene, expression values of the 1000 subjects are compared to those of the same subjects affected by the knocking out or down of the six target hubs as previously described. A gene is considered as biomarker if the median and mean differences are greater than 1e-6 and 0.05, respectively. This results in 155 biomarkers.
Once defined the biomarker list from simulated data, different synthetic dataset were generated from the two groups of 500 healthy and 500 diseased subjects: in order to consider the effect of sample size, data are partitioned into 4 sets of 10 balanced non-overlapping datasets of size 50, 20, 15 or 10 subjects per group (10 datasets for each case study), for a total of 40 simulated datasets.

2.6 Performance evaluation

Algorithms' performance was evaluated in terms of the ability to accurately classify the subjects, to provide stable lists of biomarkers and to select true biomarkers.
The Matthews correlation coefficient, MCC [77], was used as a measure of the quality of binary classifications. The MCC can be calculated directly from the

confusion matrix using the formula:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.14}$$

In this equation, TP is the number of true positive, TN the number of true negative, FP the number of false positive and FN the number of false negative subjects. MCC lives in the range $[-1, 1]$, where 1 is perfect classification, 0 an average random prediction and -1 an inverse prediction.
To evaluate the ability of the different methods to provide stable lists of biomarkers, the algebraic stability indicator derived by Canberra distance was used [62]. In particular, given two ordered lists T_1 and T_2 of p ranked features, the Canberra distance between them is defined as:

$$Ca\,(T_1, T_2) = \sum_{i=1}^{p} \frac{|\tau_1(i) - \tau_2(i)|}{\tau_1(i) + \tau_2(i)} \tag{2.15}$$

where $\tau_1(i)$ and $\tau_2(i)$ indicate the rank, *i.e.* the position, of feature i in the ordered lists T_1 and T_2, respectively. The stability indicator for a given set of lists was computed as the mean of the Canberra distances between pairs of lists in the set, normalized by its expected value on the whole permutation group on p features: the obtained value ranges then between 0 (maximal stability) and 1.4 (maximal instability), with 1 as the case of randomly generated lists.
A different extension based on quotients of permutation groups allowed comparing lists T_1 and T_2 of different length l_1, l_2:

$$Ca\,(T_1, T_2) = \frac{1}{(p - l_1)! \cdot (p - l_2)!} \sum_{\Gamma_1 \in S_1} \sum_{\Gamma_2 \in S_2} Ca\,(\Gamma_1, \Gamma_2) \tag{2.16}$$

where p is the total number of analyzed features and Γ_j (j=1,2) belong to the set S_j of all the lists having the first l_j features ordered as in T_j and the remaining $(p-l_j)$ elements ordered in all the $(p-l_j)!$ possible combinations. This is called the complete version of the partial lists distance: neglecting its component depending only on the discarded features, a different measure (called core distance) can be obtained, better tailored to highlight variations on partial short lists [78]. Full statements and proofs of the mathematical properties of the Canberra distance can be found in [79].
The ability to select the true biomarkers was evaluated in term of precision (number of true positives divided by the number of selected features) obtained by the different methods according to their choice of the optimal number of features. The area under the precision *vs.* recall (number of true positives divided by the

number of true biomarkers) curve was also considered to outline the ability of the different methods to rank the features, a task related with the ability to select the true biomarkers.
Finally, statistical significance of the comparison between each method and its bootstrap variant was assessed using Wilcoxon signed ranks test with significance level α equal to 0.05. Differences among the four multivariate methods in their bootstrap variant were assessed using Friedman test (α=0.05), followed, if significant, by Wilcoxon signed ranks test to examine between which methods the differences actually occur, with a significance level α equal to 0.05/6=0.0083 to correct for multiple testing. SAM was compared with the other eight methods using Wilcoxon signed ranks test with a significance level α equal to 0.05/8=0.00625 to correct for multiple testing.

2.7 Results on simulated data

The nine different biomarker discovery methods for binary classification and feature weighting and ranking were tested on the 40 simulated datasets of size 50, 20, 15 and 10 subjects per group, evaluating how sample size and heterogeneity characterizing simulated data affect these methods in terms of classification accuracy, stability of biomarker lists and precision of feature selection.

2.7.1 Classification accuracy

Table 2.1 presents the average MCC obtained on simulated data. All bootstrap classification methods perform equally well (Friedman test p-values always above 0.15 for every sample size) in terms of classification accuracy. In particular, bootstrap approach improves classification accuracy: with 50 subjects per group LSVM_B and IRSVM_B perform better than their standard versions (p-value equal to 0.019 and 0.007, respectively); with 20 subjects per group GSVM_B and SRDA_B perform better than their standard versions (p-value equal to 0.030 and 0.025, respectively); with 15 subjects per group LSVM_B, GSVM_B and SRDA_B perform better than their standard versions (p-value equal to 0.031, 0.031 and 0.016, respectively). SAM was excluded from this part of the analysis.

2.7.2 Feature stability

The ability of the various methods to select the same features across different datasets is depicted in Figure 2.4, where the boxplots of the core Canberra

	50	**20**	**15**	**10**
LSVM	0.73	0.69	0.73	0.70
	(0.62, 0.82)	(0.51, 0.93)	(0.60, 0.88)	(0.60, 0.82)
LSVM_B	0.77	0.74	0.80	0.73
	(0.65, 0.87)	(0.54, 0.95)	(0.68, 0.94)	(0.64, 0.83)
GSVM	0.78	0.76	0.81	0.73
	(0.70, 0.87)	(0.62, 0.91)	(0.72, 0.89)	(0.66, 0.80)
GSVM_B	0.80	0.81	0.83	0.71
	(0.65, 0.92)	(0.62, 0.95)	(0.66, 0.94)	(0.64, 0.86)
SRDA	0.75	0.72	0.74	0.69
	(0.66, 0.84)	(0.61, 0.93)	(0.61, 0.87)	(0.60, 0.80)
SRDA_B	0.77	0.74	0.75	0.73
	(0.67, 0.85)	(0.59, 0.96)	(0.60, 0.94)	(0.61, 0.83)
IRSVM	0.77	0.83	0.77	0.65
	(0.66, 0.84)	(0.61, 0.94)	(0.67, 0.85)	(0.60, 0.80)
IRSVM_B	0.81	0.72	0.80	0.69
	(0.67, 0.92)	(0.50, 0.95)	(0.64, 0.94)	(0.51, 0.86)

Table 2.1: **MCC corresponding to the optimal number of features obtained using different methods - simulated data.** Average MCC obtained when 50, 20, 15 or 10 subjects per group are available. Range of values is indicated in parenthesis.

distance (Equation 2.16) of the lists of selected features are shown. The distance between the ranked lists increases for all the methods when the number of subjects per group decreases.

The bootstrap resampling schema leads to an improvement in list stability, statistically significant when sample size decreases. In particular, differences are statistically significant for LSVM, SRDA and IRSVM with 20 subjects per group (p-value always lower than 0.036), for LSVM, GSVM and IRSVM with 15 subjects per group (p-value always lower than 0.033), for all methods with 10 subjects per group (p-value always lower than 0.001). Among bootstrap approaches, IRSVM_B is the best performing method in terms of list stability, when 20 subjects per group are available; LSVM_B performs as IRSVM_B in the case of 15 subjects per group; GSVM_B performs as IRSVM_B in the case of 10 subjects per group (Friedman test gave p-value lower than 10-11 for sample size 20, 15, 10 and Wilcoxon signed ranks test gave p-value lower than 0.001 for every significant pairwise comparison).

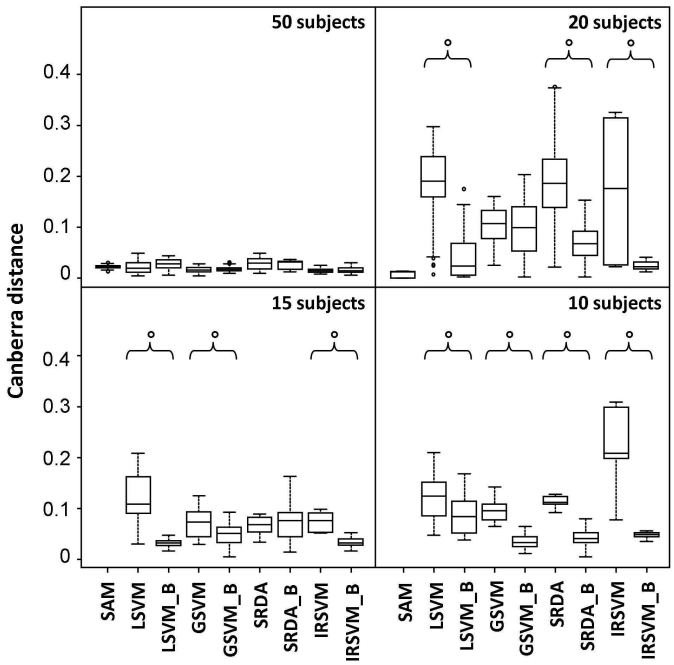

Figure 2.4: **Evaluation of feature stability on simulated data.** Boxplots of the core Canberra distance between lists of selected features obtained using different methods when 50, 20, 15 or 10 subjects per group are available. A dot highlights the significant differences between pair of bootstrap and non-bootstrap approaches (p-value lower than 0.05, Wilcoxon test).

In the case of 50 subjects per group, SAM shows list stability comparable to the one obtained by the other methods. With 20 subjects per group, SAM is as good as IRSVM_B; however, results are limited to the four datasets for which SAM was able to select features below the 0.05 FDR threshold.

2.7.3 Feature selection

Figure 2.5 shows boxplots of precision, obtained by the different methods according to their choice of the optimal number of features. Feature selection results show that bootstrap resampling schema leads to an improvement in terms of precision, statistically significant when the sample size decreases. In particular, with 20, 15 and 10 subjects per group, bootstrap improves precision of 1.5, 1.4 and 2 fold change, respectively (average improvement across the four different classification methods). Differences between bootstrap and non-bootstrap approach are statistically significant (p-value lower than 0.05, Wilcoxon signed ranks test) for LSVM and GSVM with 20 subjects per group, for LSVM and SRDA with 15

subjects per group, for all methods but LSVM with 10 subjects per group. There are no appreciable differences among different bootstrap methods in terms of precision (Friedman test p-value always above 0.05 for every sample size). In Figure 2.5, the interquartile range of the number of selected features is also reported. Interestingly, with less than 50 subjects per group, the bootstrap approaches have the tendency to select a lower number of features.

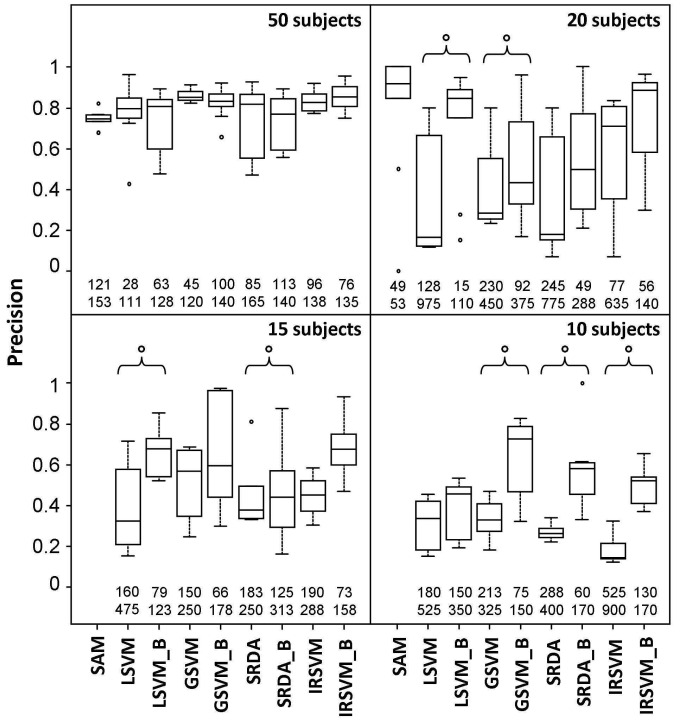

Figure 2.5: **Precision of feature selection on simulated data.** Boxplots of precision corresponding to the optimal number of features chosen by different methods when 50, 20, 15 or 10 subjects per group are available. A dot highlights the significant differences between pair of bootstrap and non-bootstrap approaches (p-value lower than 0.05, Wilcoxon test). The median number of selected features is also reported below each boxplot.

In the case of 50 subjects per group, SAM detects differentially expressed features with average precision comparable to that obtained by the other methods, but GSVM, IRSVM and IRSVM_B, which perform statistically significantly better than SAM (p-value equal to 0.002, 0.006, 0.006 respectively, Wilcoxon signed ranks test). With 20 subjects per group, SAM is not able to select any gene with FDR lower than 0.05 in six datasets, whereas in the remaining four, it selects in average 50 features with high precision (0.85 in average). In these latter cases

SAM performs statistically significantly better than LSVM (p-value=0.004) and SRDA (p-value=0.006), *i.e.* two methods without the bootstrap approach. Finally, with less than 20 subject per group, SAM is not able to select any gene in any of the dataset with FDR lower than 0.05; thus no result could be reported in these latter two cases.

A slightly different task, although related to feature selection, is feature ranking. In principle, a method could rank features properly, but fail to select the optimal number of features. Areas under the precision *vs.* recall curves (AUC) obtained by ranking features (Figure 2.6) show appreciable differences between methods. Bootstrap methods perform better than their standard variants for datasets of size 50, 20 and 15, for all methods (p-value always below 0.005) but GSVM. For datasets of size 10, only SRDA_B improves with respect to SRDA (p-value=0.01). With datasets of 50, 20 and 15 subjects per class, IRSVM_B is the best performing algorithms (Friedman test gave p-value lower than 0.004 for every sample size and Wilcoxon signed ranks test gave p-value lower than 0.003 for every comparison be-tween IRSVM_B and the other bootstrap methods). With 10 subjects per group, all multivariate methods show AUC below 0.5, without statistically significant differences among them.

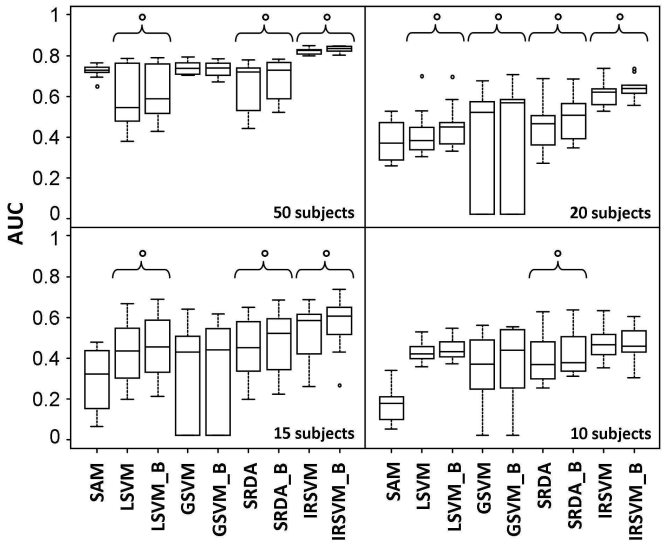

Figure 2.6: **Evaluation of feature ranking on simulated data.** Boxplots of area under the precision *vs.* recall curves obtained by ranking features according to the different methods, when 50, 20, 15 or 10 subjects per group are available. A dot highlights the significant differences between pair of bootstrap and non-bootstrap approaches (p-value lower than 0.05, Wilcoxon test).

With 50 and 20 subjects per group, a simple univariate test such as SAM is able to rank differentially expressed features with performance comparable to multivariate methods such as LSVM, GSVM, SRDA and their bootstrap versions, but not to IRSVM and IRSVM_B that perform better (p-values equal to 0.002 for both tests). However, when the number of subjects is lower than 20, SAM performance in feature ranking dramatically drops with respect to classification based methods (p-value lower than 0.002 for all comparisons but GSVM and GSVM_B). This behavior is consistent with the inability of SAM to select any feature with 15 and 10 subjects per group.

2.8 Application to Real datasets

In order to validate results obtained on simulated data comparing classification methods in terms of classification accuracy and consistency of lists of candidate biomarkers on real case studies, three publicly available microarray datasets monitoring breast cancer patients with positive and negative estrogen receptor status were used; biomarker lists from the three datasets as well as sets of sub-lists of different sample size obtained from each dataset were compared.
Breast cancer microarray studies were collected from Gene Expression Omnibus repository (GEO) with accession numbers: GSE2990 [80], GSE3494 [81] and GSE7390 [82]. Datasets were all hybridized using Affymetrix U133 Genechips^{TM} (HG-U133A). Samples that have known estrogen-receptor (ER) status were selected so to have balanced groups (ER+ and ER-), homogeneous with respect to characteristics such as age, tumor size and histological grade. ER status was chosen as case study because it is always assessed in breast biopsies, therefore it is very often present among the clinical/pathological information given with the datasets. Moreover, the assessment of the ER status is important to divide breast cancer into molecular classes and to treat cancer with the hormone blocking therapy [83]. Since there are subgroups of samples belonging to multiple datasets, redundant subjects were removed. The resulting datasets are characterized by 22207 features (probe sets) and 66 subjects for GSE2990 (33 ER+, 33 ER-), 50 subjects for GSE3494 (25 ER+, 25 ER-) and 92 subjects for GSE7390 (46 ER+, 46 ER-). Comparison among the three datasets allowed assessing list stability in a real case study. To assess list stability within dataset, thus not accounting for experimental setup variability, and to compare the effect of sample size with simulated data, 20 subjects per ER status were repeatedly sampled from datasets GSE2990 and GSE7390 to set up smaller balanced datasets (10 datasets for each

case study). Gene expression intensity signal was derived and normalized independently for each dataset using the robust multiarray average (RMA) algorithm [84]. Probe sets related to the estrogen receptor (ESR1) were removed from all datasets, since ESR1 is the gene more directly associated with ER status and can mask other potential descriptors of the underlying pathophysiology [85].
In terms of classification accuracy, the MCC obtained using different methods on real datasets is shown in Table 2.2. Results on dataset GSE3494 are not shown since none of the different methods gave good accuracy (MCC always below 0.4). On the other two datasets, results confirmed those obtained by simulated data. The first two columns report the MCC for GSE2990 and GSE7390, respectively, when 20 subjects per group are repeatedly sampled from each dataset. The third and fourth columns of Table 2.2 report the MCC obtained using the complete datasets GSE2990 and GSE7390. Results are comparable to those obtained using simulated data. Bootstrap approach improves classification accuracy on dataset 7390 for all methods (p-value equal to 0.02, 0.04, 0.001, 0.03 for LSVM_B, GSVM_B, SRDA_B and IRSVM_B, respectively, with respect to their standard version), whereas, with dataset 2990, the differences between bootstrap and standard approaches are not statistically significant. As observed with simulated data, all bootstrap classification methods perform equally well (Friedman test p-values always above 0.06 on both the datasets). In terms of stability, bootstrap resampling schema leads to an appreciable improvement both when the complete datasets GSE2990 and GSE7390 are compared and when 20 subjects per group are repeatedly sampled from each dataset (Figure 2.7).
Differences between bootstrap and standard approach are statistically significant for every method (p-value always lower than 0.002) with dataset GSE2990 and for LSVM and GSVM with dataset GSE7390. Observing the interquartile range of the number of selected features reported in Figure 2.7, it is confirmed the tendency of the bootstrap approaches to select a lower number of features.
SAM performance is poor: when 20 subjects per group are repeatedly sampled from each dataset the core Canberra distance between lists of biomarkers ranges between 0.04 and 0.37 (average 0.27) for GSE2990 and between 0.13 and 0.31 (average 0.22) for GSE7390; on the other hand, between the complete datasets (GSE2990 vs. GSE7390) the core Canberra distance is equal to 0.63. SAM results are not shown in Figure 2.7 to avoid masking the differences among the other methods.

	GSE2990 20 subjects	**GSE7390 20 subjects**	**GSE2990**	**GSE7390**
LSVM	0.64 (0.61, 0.69)	0.77 (0.61, 0.90)	0.6	0.79
LSVM_B	0.65 (0.51, 0.77)	0.81 (0.58, 0.91)	0.68	0.81
GSVM	0.62 (0.59, 0.64)	0.73 (0.60, 0.83)	0.59	0.74
GSVM_B	0.65 (0.60, 0.71)	0.78 (0.61, 0.91)	0.61	0.77
SRDA	0.63 (0.61, 0.66)	0.74 (0.62, 0.85)	0.5	0.78
SRDA_B	0.67 (0.61, 0.78)	0.83 (0.66, 0.90)	0.67	0.77
IRSVM	0.62 (0.47, 0.69)	0.80 (0.65, 0.91)	0.6	0.78
IRSVM_B	0.67 (0.58, 0.82)	0.82 (0.62, 0.91)	0.67	0.81

Table 2.2: **MCC corresponding to the optimal number of features obtained using different methods real data.** Average MCC obtained when 20 subjects per group are available, sampled from datasets GSE2990 and GSE7390 MCC (first and second columns, respectively; range of values is indicated in parenthesis), and obtained on the complete datasets GSE2990 and GSE7390 (third and fourth columns, respectively).

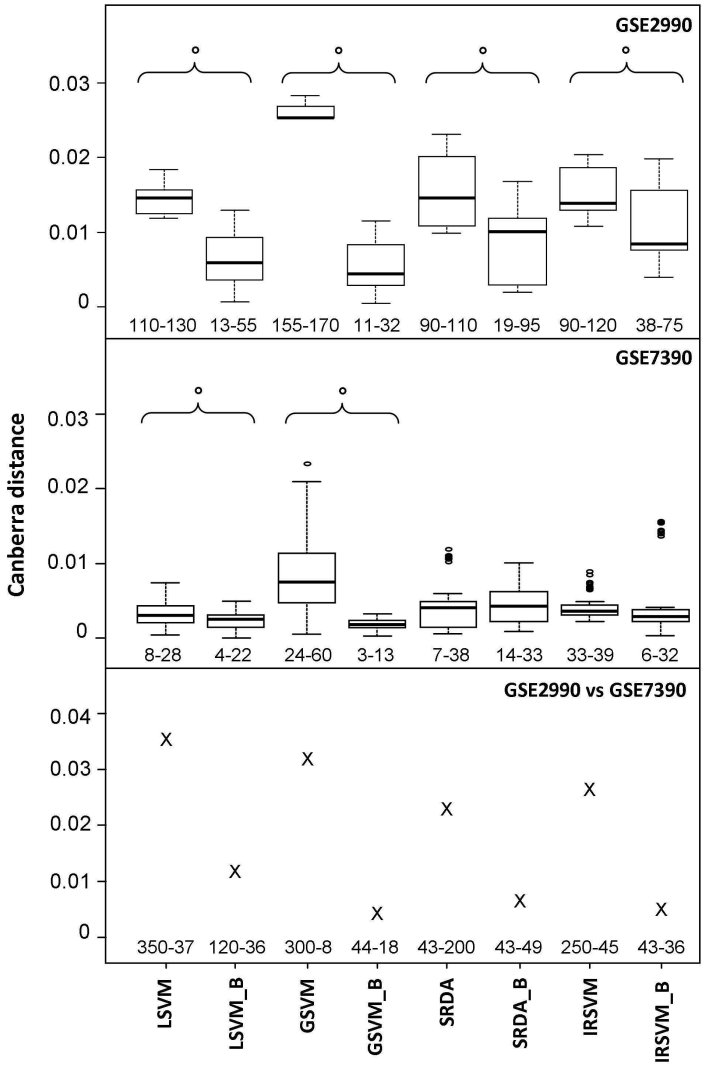

Figure 2.7: **Evaluation of feature stability on real data.** Boxplots of the core Canberra distance between lists of selected features provided by different classification methods when 20 subjects per group are repeatedly sampled from GSE2990 (upper panel) and GSE7390 (middle panel) datasets. A dot highlights the significant differences between pair of bootstrap and non-bootstrap approaches (p-value lower than 0.05, Wilcoxon test). The interquartile range of the number of selected features is reported below each boxplot. The core Canberra distances between lists of biomarkers provided by different methods on the complete GSE2990 *vs.* GSE7390 datasets are shown in the lower panel together with the number of selected features in each dataset.

2.9 Discussion

For complex diseases such as cancer, high throughput analysis carried out in different research centers may exhibit poor reproducibility, with limited overlap or reduced statistical significance. The results of the MAQC-II study address in a comprehensive analysis this issue on real datasets by comparing methods and procedures between data analysis teams [44]. Here, the effect of the intrinsic complexity of this task has been further explored.
A first contribution of this work is the comparison of different classification methods applied on real microarray datasets, in terms of consistency of lists of candidate biomarkers and classification accuracy. A second contribution is the generation of a simulated dataset to extensively assess average method performance on a large number of studies and experimental conditions accounting for the low number of samples available and the heterogeneity affecting microarray studies, and to evaluate precision and feature ranking performance on a benchmark with known biomarkers. Heterogeneity of samples in each group is obtained by simulating both intrinsic variability of the population and heterogeneity of the disease. Despite its simplicity with respect to real systems, the simulator provides a versatile test bed to assess a wide spectrum of methodologies.
Results on simulated data show that when some tens of subjects are available per group, performance of different methods are comparable. However, when available subjects are equal or lower than 20, bootstrap resampling schema leads to an improvement in list stability and the precision of the selected features. Bootstrap approach slightly improves also classification accuracy when 50, 20 or 15 subjects per group are available. Among the different methods here considered, IRSVM_B provides the best combination of feature ranking and biomarker stability; moreover, it reaches the best average performance also in terms of classification accuracy. In the case of 50 subjects per group, a simple univariate test such as SAM shows performance comparable to that obtained by the other methods, thus confirming results obtained in [64], where in the four used datasets the lowest number of subject per group is always above or equal to 40. However, with 20 subjects per group, SAM performance strongly depends on the dataset: on the simulated data, for example, SAM is not able to select any gene with FDR lower than 0.05 in six datasets, whereas in the remaining four, it selects in average 50 features with high precision (0.85 in average) and stability comparable to the one obtained using IRSVM_B, although this latter outperforms SAM in feature ranking. Finally, with less than 20 subjects per group, SAM performance dramatically drops

with respect to classification based methods. With real data, only list stability and classification accuracy can be assessed. In both cases, results of classification methods tightly resemble those obtained with simulated data.
In conclusion, the analysis confirms the MAQC-II indication that comparably good classification accuracy can be reached by different methods on the same task, if a valid Data Analysis Plan is adopted [86]. Furthermore, the study highlights a systematic improvement due to bootstrap approach in selecting features with a high degree of precision and stability. Overall, the crucial factor affecting list stability seems to be that the classification task is under constrained. When additional information is present on the relationships between genes, this information could be used to improve the stability with respect to the features of the classifiers. The basic idea of this strategy would be to take into account the complex gene relationships, instead of considering genes as independent features. In the following chapters, the effect of the use of different biological information from genomic databases in the learning process is explored, by integrating different prior knowledge like functional annotations, protein-protein interactions, and expression correlation among genes.

Chapter 3

Biological knowledge in genomic databases

For many years research in molecular biology has been focused on the analysis of relevant components (genes, proteins, metabolites) of a particular cellular process in isolation. By this approach many genes have successfully been characterized and functionally annotated, but biological systems are complex and their characteristics are the result of a highly interwoven interaction network developing through time and space. Fundamental biological processes of living systems, *e.g.* assimilation of nutrients and the perception of environmental signals, are the result of the interplay of different biochemical reactions, thus the understanding of these processes is essential. However, this requires an approach that takes into account both interactions at the molecular level as well as physiological functions that are characteristics of the whole organism. From this point of view, systems approaches are becoming increasingly important, since the understanding of multigenic and complex diseases cannot be related to a single gene or component but to multiple pathways and the interplay between different genes and the environment.

The development of systems approaches requires a lot of information on different aspects of the system. Data typically arise from several levels of cellular information. The most important resource for such information is the scientific literature and human expertise curated one in public databases. However, as more and more genomes have been sequenced and annotated, and protein and gene interaction data have been accumulating, recent years have seen an explosion in the amount of available biological data. Biological databases have become an essential instrument for managing these data and for making them accessible. Data are collected from scientific experiments, published literature, high-

throughput experiment technology, and computational analyses. Depending on the data that they contain, these databases fulfill different functions. The collected information derives from different research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. This information is often described as semi-structured data, and can be represented as tables, key delimited records, and XML structures. An example of data representation is the *ontology*, which provides both a vocabulary representing knowledge as a set of concepts within a domain and a graph of the relationships between these concepts. The ontology structure reflects the current representation of knowledge as well as serving as a guide for organizing new data. A popular biological ontology is the *Gene Ontology*, where genes and proteins can be annotated to biological concepts at varying levels of the graph depending on the amount and completeness of available information. Other structured data representations are gene and protein networks, where each relation between two biological entities (*i.e.* genes or their molecular products) can represent a specific biochemical interaction or a regulatory mechanism which can occur at different levels of a cellular process. Depending on the nature of the relations represented in these biological networks, there are different databases.
In the following, an overview of existing database for the biological information on processes, functions and interactions is provided, in particular focusing on Gene Ontology and protein-protein interactions databases.

3.1 Gene Ontology database

The Gene Ontology (GO) project [31] is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The project began as a consortium among three model organism databases, FlyBase (Drosophila) [87], the Saccharomyces Genome Database (SGD) [88] and the Mouse Genome Database (MGD) [89], in 1998. The GO project has developed three structured controlled vocabularies (categories) that describe gene products in terms of their associated biological processes, molecular functions and cellular components in a species-independent manner. In particular:

- *Biological Process* deals with sets of molecular events, *i.e.* chemical or physical transformation, with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms;

- *Molecular Function* collects the elemental activities of a gene product at

the molecular level, such as binding or catalysis, describing only what is done without specifying where or when the event actually occurs;

- *Cellular Component* represents the parts of a cell or its extracellular environment where a gene product is active.

Each GO term within the ontology has a term name (which may be a word or string of words), a unique alphanumeric identifier, a definition with cited sources and a namespace indicating the domain to which it belongs. Terms may also have synonyms, which are classified as being exactly equivalent to the term name, broader, narrower, or related.
To each GO term, a set of annotation (gene products, proteins, etc.) is associated, where each element can be distinguished by an identification code (*e.g.* EntrezID, UniProtID, etc.). All annotations are characterized by an *evidence code*, which comes from the Evidence Code Ontology, a controlled vocabulary of codes covering both manual and automated annotation methods. The following sections illustrate in more detail the structure and the properties of annotations of GO database.

3.1.1 Graph structure

GO terms are organized in a directed acyclic graph (DAG) in which each node corresponds to a GO term and each edge to a relationship between two GO terms. An example of the GO structure is shown in Figure 3.1, where the node related to GO term *regulation of cell projection assembly* is displayed with all available paths to its root term. Each node may have multiple parents: nodes farther from the root (high level nodes) correspond to more specialized terms, nodes closer to the root (low level nodes) to less specialized ones. The main relationships between GO terms are *is-a* and *part-of* relationships. The is-a relationship connects a subtype to its more general counterpart, but it does not mean "is an instance of": GO, like most ontologies, does not use instances, and the terms in GO represent a class of entities or phenomena, rather than specific manifestations. The is-a relation is transitive, which means that if A is-a B, and B is-a C, then it is possible to infer that A is-a C. The relation part-of is used to represent part-whole relationships in GO; part-of has a specific meaning in GO, and a part-of relation would only be added between A and B if B is necessarily part of A: wherever B exists, it is as part of A, and the presence of the B implies the presence of A. However, given the occurrence of A, B might not exist. Like is-a, part-of is transitive.

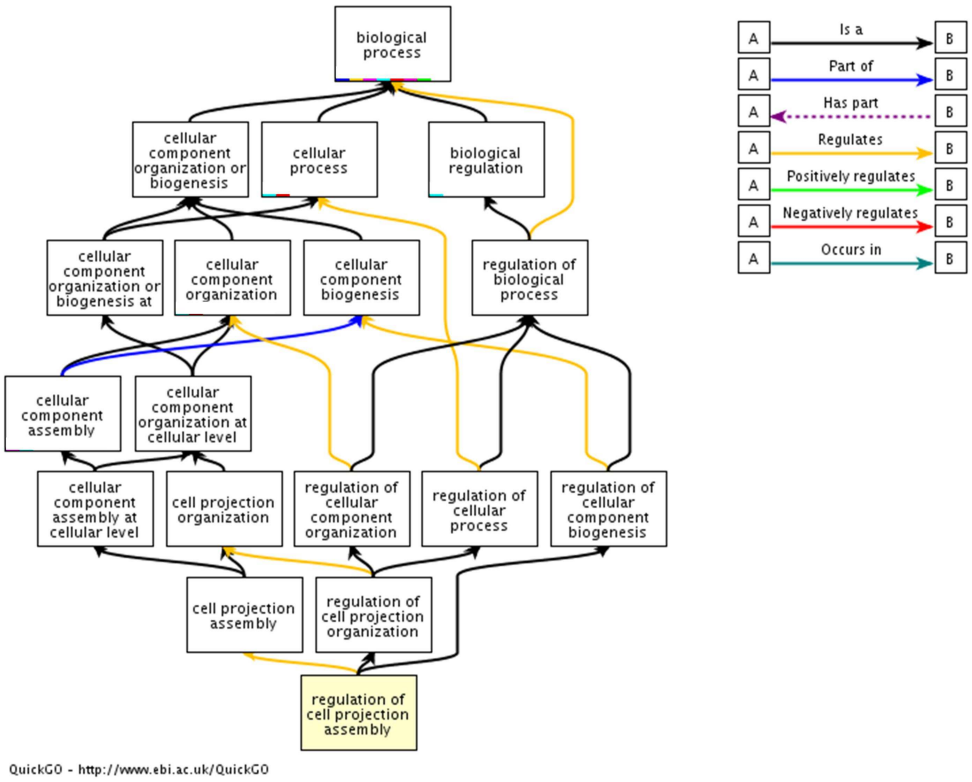

Figure 3.1: **Example of the GO structure.** The structure of the GO is illustrated on the paths of term *regulation of cell projection assembly* to its root term.

The original use of the part-of relationship between regulatory processes and the processes that they regulate did not provide enough specificity to allow users to perform queries that distinguish gene products that play a regulatory role versus a direct role in a biological process. Recently, new relationships were introduced in the ontology: *regulates*, *positively-regulates*, and *negatively-regulates* relationships between regulatory terms and their regulated parents. The three regulates relationships allow GO to correctly represent important areas of biology where one process affects the manifestation of another process but may not be a part of that process itself. For example, *regulation of transcription* is not a part of *transcription*, but lies outside of the transcription process and controls how it unfolds. The "regulates" relations in GO are used specifically to mean necessarily-regulates, that is: if B regulates A, then whenever B is present, it always regulates A, but A may not always be regulated by B.

In the last years, GO curators also introduced the *has-part* relationship. It represents a part-whole relationship from the perspective of the parent, and is thus the

logical complement to the part-of relationship. In GO, the relationship A has-part B means that A necessarily (always) has B as a part; *i.e.* if A exists then B also exists as a part of A. If A does not exist, B may or may not exist. For example, *cell envelope* has-part *plasma membrane* means that a cell envelope always has a plasma membrane as a part but a plasma membrane may exist without being a part of a cell envelope.
The GO undergoes frequent revisions to add new relationships and terms or remove obsolete ones. If a term is deleted from the ontology, the identifier for the term stays valid, but is labelled as *obsolete* and all relationships to the term are removed. Changes to the relationships do not affect annotations, because annotations always refer to specific terms, not their location within the GO. Recently, also relationships between the three categories were introduced, thus eliminating orthogonality among the information in these three domains. Currently, there are two concurrent versions of the GO, the filtered and the full GO. The main difference is that the filtered GO does not contain any has-part or inter-ontology relationships. In this thesis, the filtered version of GO was used.

3.1.2 Functional annotations

A GO annotation associates a gene with terms in the ontology and can be generated either by a curator or automatically through predictive methods. Genes are associated with as many terms as appropriate as well as with the most specific terms available to reflect what is currently known about a gene. When a gene is annotated to a term, associations between the gene and the terms' parents are implicitly inferred. Because GO annotations to a term inherit all the properties of the ancestors of those terms, every path from any term back to its root must be biologically accurate or the ontology must be revised. Thus, a gene annotated to a specific GO term can be retrieved not only with that term, but also with all of its parent terms, increasing flexibility and power when searching for and making inferences about genes, but at the same time introducing strong dependencies among the GO terms and redundancy of the information. Each annotation in the GO has a source and a database entry attributed to it. The source can be a literature reference, a database reference or computational evidence. One of the most important attributes of an annotation is the evidence code. These evidence codes are divided into four categories (Figure 3.2). For example, Traceable Author Statement (TAS) means a curator has read a published scientific paper and the metadata for that annotation bears a citation to that paper; Inferred from Sequence Similarity (ISS) means a human curator has reviewed the output from

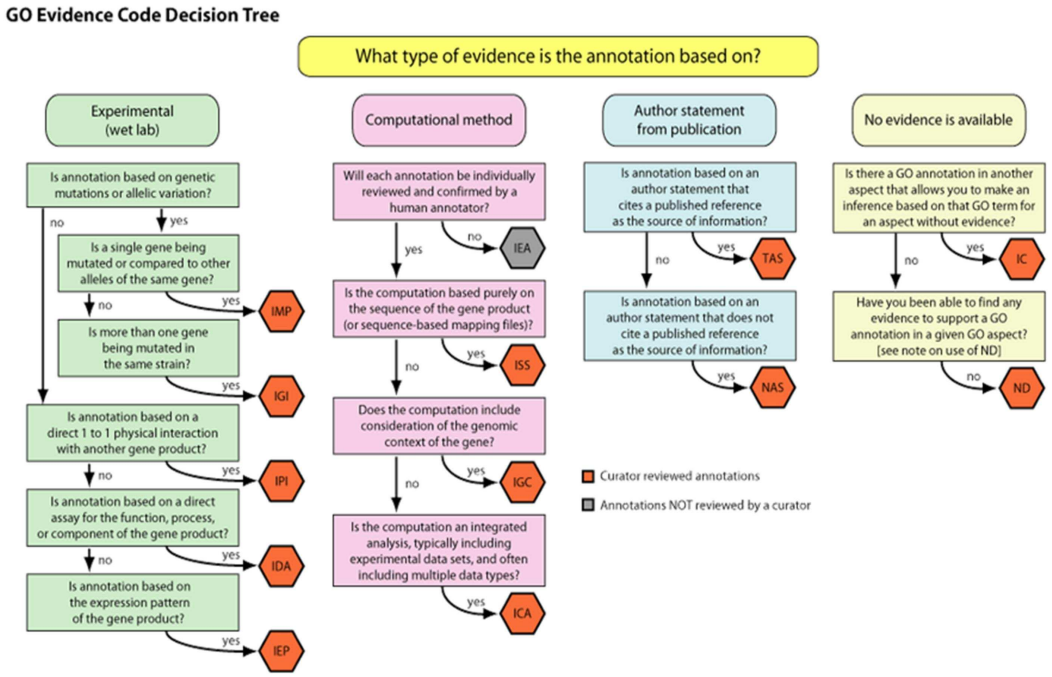

Figure 3.2: **A decision tree for deciding which evidence code to use.** Figure adapted from `http://www.geneontology.org/GO.evidence.tree.shtml`

a sequence similarity search and verified that it is biologically meaningful. The ND evidence code indicates that the function is currently unknown (*i.e.* that no characterization of the gene is currently available). Annotations from automated processes (for example, remapping annotations created using another annotation vocabulary) are given the code Inferred from Electronic Annotation (IEA). The most reliable annotations are those inferred directly from experimental evidence, in particular from direct assay (IDA). GO annotations are continually updated to reflect current knowledge, to correct errors and to improve logical consistency. The GO ontology is updated daily and most of the annotation files are released weekly. Although annotations are robust to changes in the ontology because they are made to the definition of the term and not to the term name or its position in the graph, it is important to use the latest versions of the ontology and annotations.

3.2 Pathway databases

A pathway is a set of interactions, or functional relationships, between the physical and/or genetic components of the cell which operate in concert to carry out a biological process [90]. Pathway databases put protein interactions into a biological context by creating pathways to describe biological processes, representing a connected sequence of biochemical reactions. These pathway databases facilitate a variety of analyses and simulation techniques that can enrich the understanding of cellular systems. The Pathway Resource List [91] currently provides an overview of more than 190 web-accessible biological pathway and network databases, including information on metabolic pathways, signaling pathways, gene regulatory networks, genetic interactions and protein-protein interactions. Metabolic pathway databases generally store a series of biochemical reactions, focusing mainly on the chemical modifications made to the small molecule substrates of enzymes. Signaling pathway databases generally collect sets of molecular interactions and chemical modifications, propagating information from one part or sub-process of the cell to another, often via a series of protein covalent modifications, such as protein phosphorylation. Dysregulation of biological processes by aberrant signaling pathways causes many common diseases, such as cancer and diabetes. Gene regulation network databases capture transcription factors and the genes they regulate; these databases share features with both signaling and protein interaction databases, as they collect protein-DNA interactions and regulatory (activation and inhibition) events. Genetic pathway databases are composed of genetic interactions which occur when two mutations have a combined phenotypic effect that is not simply the sum of the effects caused by either mutation alone. Finally, protein-protein interaction databases, which will presented in more detail in the following, mainly store pairwise interactions or complexes between proteins and sometimes other molecular interaction types. Unlike metabolic and signaling pathway data, which are generated primarily by traditional small-scale experimental techniques, large amounts of protein interactions (protein-protein, protein-DNA, etc.) are generated by various large-scale experimental methods.

All these pathway databases are interesting for modelling approaches, since they offer a straightforward way of building network topologies by the annotated reaction systems [91], however they provide also an integrated topological representations of functional knowledge of the different components of a biological system which allow using this information as a prior knowledge for different genomic and

proteomic analyses. Examples of these databases are the Kyoto Encyclopedia of Genes and Genomes (KEGG) [43], Reactome [92] and the Pathway Interaction Database (PID) [93], which contain metabolic reactions and several signal transduction pathways. KEGG is a reference knowledgebase offering information about genes and proteins, biochemical compounds, reactions and pathways. It provides 317 reference pathways that are linked to genes and reactions of several organisms. Reactome uses a very precise specification (ontology) of components and interactions that comprises details on stoichiometry, localization, references to external databases, etc. This covers also processes like complex formation events or translocations of molecules. For signalling events, PID is a growing collection of human signaling and regulatory pathways curated from peer-reviewed literature and stored in a computable format. Gene regulation processes and gene regulatory networks are not yet covered in as much detail as metabolic processes or signalling. However, there are databases that store information on transcription factor binding sites such as TRED [94], TRANSFAC [45] and JASPAR [46]. An inherent aspect of the pathway concept is protein-protein interaction represented by several databases such as IntAct [95] or database of interacting proteins (DIP) [96], which accomplish a comprehensive knowledge of the protein interactomes from studies on proteome-wide physical connections between protein pairs measured by efficient large-scale technologies. This last type of pathway-related information will be presented in detail in the following section.

3.3 Protein-protein interactions

In recent years, given an explosive development of high-throughput experimental technologies, the number of reported protein-protein interactions (PPIs) has increased substantially. A protein-protein interaction (PPI) network is commonly viewed as an unweighted, undirected graph. Each node in the graph represents a protein and an edge between a pair of nodes indicates that these proteins have been observed to interact physically (Figure 3.3). In an attempt to understand and describe the PPI connectivities, a number of models for generating edges in some probabilistic sense, have been proposed and tested against observed networks [97, 98]. Many works have focused on matching degree distributions and recovering a scale-free law [73, 99], although whether PPI networks are really scale-free is still the subject of debate [97, 100].

The primary resources for PPI data are individual scientific publications. To make this information more readily available, a number of publicly available databases

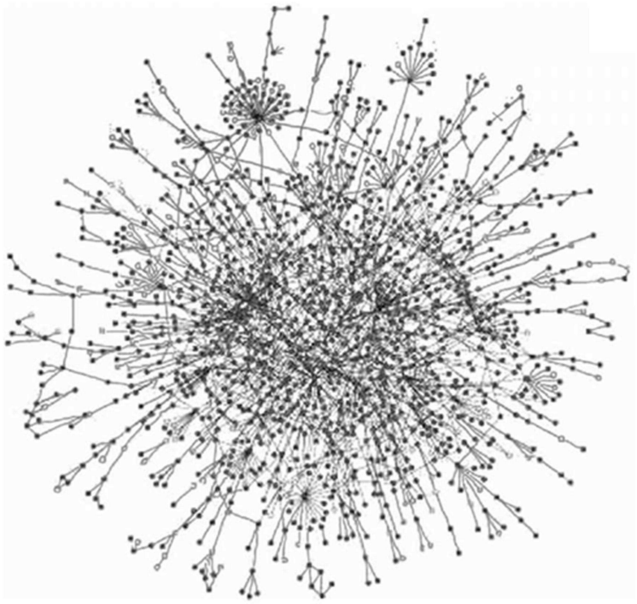

Figure 3.3: **A map of protein-protein interactions for 1870 yeast proteins.** Figure from Jeong *et al.*, 2001 [3]

have set out to collect and store protein-protein interaction data, providing a global view of protein partners and protein memberships in molecular complexes. These usually reference the original publication and the experimental method that determined every individual interaction. In order to uniform this knowledge, the International Molecular Exchange (IMEx; `http://imex.sourceforge.net/`) consortium was formed. The primary PPI databases are DIP [96], IntAct [95], and MINT [101], which are the core founders of IMEx. Another complete PPI database is the Human Protein Reference Database (HPRD) [47], which focuses entirely on human proteins, providing not only information on protein interactions, but also a variety of protein-specific information, such as post-translational modifications, disease associations and enzyme-substrate relationships.
Focusing on human proteins, the Human Protein Reference Database (HPRD) contains manually curated scientific information pertaining to the biology of most human proteins and is the database that includes most human protein-protein interactions, as shown in [102]. The National Center for Biotechnology Information provides link to HPRD through its databases (*e.g.* Entrez Gene, RefSeq) pertaining to genes and proteins. Moreover, HPRD provides not only information on protein interactions, but also a variety of protein-specific information, such as post-translational modifications, disease associations and enzyme-substrate relationships. Protein annotation information was derived through manual curation

using published literature by expert biologists and through bioinformatics analyses of the protein sequence. From 10,000 protein-protein interactions (PPIs) annotated for 3000 proteins in 2003, HPRD has grown to over 36500 unique PPIs annotated for 25000 proteins.

Chapter 4

Gene Ontology based classification: improving prediction and biological interpretability

4.1 Background

All statistical methods to perform classification and feature selection allow the identification of gene lists. High-throughput genomic experiments often lead to the identification of large gene lists, which are affected by several problems related to the stability of the solutions. Moreover, these methods for finding interesting genes often do not help the interpretation of the resulting gene lists and the formulation of consistent biological hypotheses from these results poses a challenging task. Most of the methods proposed in the literature are based on a posteriori annotation of the selected features, in order to describe the main biological processes characterizing the results. Searching for sets of predefined functionally related genes (*e.g.* pathways) that are enriched in a gene list is a popular approach to solve this problem. In particular, enrichment analysis is a statistical technique to analyze and interpret large gene lists using a priori-knowledge [103]. It assesses the over- (or under-) representation of a known set of genes (*e.g.* a biological process) within the input gene list [104, 105, 106]. A statistically significant number of genes from the known set in the gene list may indicate that the biological process plays a role in the biological conditions under study. This analysis is repeated for all available known gene sets, obtaining a score (usually

a p-value) for each set. Different enrichment methods were developed [107] and many of these were applied on gene sets from Gene Ontology (GO) annotations, since they are readily accessible for many organisms and cover many genes. However, most of these approaches are threshold-dependent, because they require the user to set a threshold on the gene scoring statistic. Specifically, results may not be stable to choice of the threshold, since there is loss of information caused by treating gene scores in a binary way (they either pass the threshold or not).
A different approach is proposed by knowledge-driven classification and feature selection methods, which integrate biological knowledge into the learning process in order to identify biomarkers not as individual genes but gene sets. Lottaz and Spang [32] proposed a structured analysis of microarray data (StAM), which generates a classifier graph according to the Gene Ontology, constructs leaf node classifiers based on selected expression values from shrunken centroid classification and propagates classification results through the inner nodes to the root by a weighted sum, where the weights are related to the performance of the classifiers. A shrinkage scheme was used to shrink the weights towards zero so that a sparse representation is possible. In the proposed method the user has to specify two calibration parameter which control the performance of the classifiers and the shrinkage process. Since the final classifier is built based on the GO tree, it greatly facilitates the interpretation of a final result in terms of identified biological processes that are related to the outcome. However, only the genes annotated in the leaf nodes (*i.e.* with most detailed biological functions) are used as predictors; because of incomplete knowledge, other relevant genes that are not annotated to leaf nodes cannot be used, missing important genes and losing predictive performance of the final model.
Following the assumption that genes in the same group are more likely to function similarly, Tai and Pan [33] proposed a group penalization method that use group-specific penalty terms and associated penalty parameters to account for possibly varying degrees of relevance of the gene groups to the outcome of interest. The basic idea is to treat the genes from the same group to be equal a priori while those from different groups unequal a priori. In the proposed method, the shrinkage parameter for a group is inversely weighted such that multiple group-specific shrinkage parameters are tuned only by one regularization parameter. The weighted method penalizes less on the genes in a group with a larger mean parameter estimate: when a group contains a larger proportion of non-zero coefficients or a few large non-zero coefficients, indicating the existence of potentially useful genes in the group, the coefficients of the genes in the group will be shrunken less

likely to be zero, leading to both higher chance of identifying important genes and in general smaller biases of shrinkage estimates. This approach is based on the idea that if the genes in a pathway or from the same functional group tend to work together, it is more likely that the genes from the same group have either zero or non-zero parameters simultaneously. This assumption strongly depends on the type of information used as prior-knowledge, which often provides groups of genes characterized by different level of specificity of biological information. In this work the method was applied on KEGG pathways, but other sources of biological knowledge for gene functions or pathways, such as GO, can be also utilized in their proposed methods. However, how to take advantage of the hierarchical structure of GO annotations is not yet defined.

These last two methods are based on the assessment of classification performance in terms of prediction accuracy. More recently, the method proposed by Haury *et al*. [64] addresses also the stability issue by combining the recently proposed graph Lasso procedure [108] with a stability selection procedure, akin the "randomization and aggregation" approach described in [27], where the feature selection process is repeated on many randomly perturbed training sets by a bootstrap approach, keeping only those features that are often selected in this procedure. Reproducibility of biomarkers lists was assessed adopting a gene-centric view, however it is important to note that the instability characterizing the identified gene lists is also affected by heterogeneous genomic alterations, which possibly affect a specific set of biological process, but not the same genes in different patients [109].

Another problem related to the use of biological knowledge is represented by the redundancy of information, which is particularly problematic with gene-sets derived from hierarchical functional annotation systems, like GO, as children terms are partially redundant with their parents by definition. Existing methods used for enrichment analysis were developed to take advantage of the GO structure to reduce redundancy. GOstats [110] and elim approach [111] integrate GO graph topology in enrichment analysis: child terms (*i.e*. leaf nodes) are tested first, then parent nodes are modified in order to not include the genes belonging to their enriched children. As gene sets collections get larger and more complex, integration methods experience longer lists of results and increased redundancy between sets, thus new approaches able to manage biological information on a global scale are required.

4.2 GO based classification method

The search for predictive biomarkers from high-throughput data is hampered by heterogeneity of diseases which strongly affect the interpretability of results, however a robust approach for the identification of biomarker lists has not as much to force the reproducibility of a gene list, but rather to select coherently the underlying biological pathways or processes which are altered by the disease. The semantic organization of GO database can be exploited to build structured feature sets constraining the learning process, so to select gene sets sharing similar biological processes or molecular functions. In this way, molecular signatures are represented by lists of biological functions compromised by the disease instead of a list of independent or partially correlated biomarkers, thus increasing the interpretability of the results of classification and feature selection analyses. However, when a gene set approach is adopted to performed a knowledge-based classification, a stable representation of biological functions is necessary to obtain an efficient characterization of the disease alterations. GO is a controlled vocabulary where genes annotated with a specific node are also annotated with every ancestor of that node, thus introducing high redundancy of the information.
Starting from the approach proposed in [32], the proposed method builds a classifier for each gene set annotated to a GO term and controls the redundancy of information characterizing inner nodes by using an elimination strategy proposed in [111], in order to improve class prediction and increase the biological interpretability of the results by selecting stable biological processes and molecular functions characterized by a high information content. The method exploits the direct acyclic graph of the Gene Ontology to define different sets of genes sharing the same annotation. For each gene set, classification analysis and feature selection are based on $\ell_1\ell_2$ regularization approach which will be presented in detail in the next section, using a double optimization schema described in [112]. Starting from this classifier, the method addresses the redundancy problem affecting GO annotation using the *elim* strategy. This approach investigates the nodes in the GO graph bottom-up. It starts processing the nodes from the highest level nodes, *i.e.* the farthest from the root, which are the most specific, and then iteratively moves to nodes of a lower level. Since nodes from the same level share no edge, they can be investigated independently. The bottom-up strategy assures that for a currently investigated node all children have been scored. The level attributed to each node is defined considering the maximum path length from the root node. When a node is processed, the classifier is applied on data only considering the

features representing the genes which are annotate to the GO term representing that node. Classification performance is measured in terms of Matthews Correlation Coefficient (MCC, section 2.6). If the MCC exceeds a fixed threshold, then the genes selected by the $\ell_1\ell_2$ approach are removed from the annotations of all the ancestors of that node. In this way, the method both accounts for the redundancy of GO annotations in the ontology and preserves the specificity of the biological information associated to the selected genes, because it promotes gene selection on the most specific GO nodes in the graph. The algorithm terminates when the root node is analyzed. At the end, the method provides a list of the selected GO terms, which could be ranked according to the MCC obtained in the test set applied to each node.

To each GO term, a classification model, specific for the genes selected by the $\ell_1\ell_2$ classifier, is provided. In order to classify new subjects, the method uses the predictive models obtained on selected gene sets to build a multiple classifier, adopting a majority vote strategy to assembling results. In detail, for each node i the $\ell_1\ell_2$ classifier provides the estimated weight vector β_i, indicative of the predictive power of each feature. Thus, the prediction of new subjects performed by the model built on the node i, $\hat{\mathbf{y}}_i$, can be obtained as $\hat{\mathbf{y}}_i = \beta_i \times D_i$, where D_i represents the dataset restricted to the genes annotated to the node i. Since the $\ell_1\ell_2$ regularization approach allows zero-values in the weight vector β_i performing feature selection, the prediction is based only on the most discriminating genes selected by the $\ell_1\ell_2$ model. Note that, considering the subject j, $-1 \leq \hat{\mathbf{y}}_i \leq 1$, the absolute value of $\hat{\mathbf{y}}_i(j)$ can be interpreted as an indicator of the reliability of prediction result. The mean prediction $\mathbf{y}_{mean}$ obtained over the set of the selected GO nodes is then obtained by:

$$\mathbf{y}_{mean} = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbf{y}_i \tag{4.1}$$

with N_s indicating the number of selected GO terms. The final prediction rule is defined by the following formulation:

$$\mathbf{y} = \begin{cases} 1 & \text{if } \mathbf{y}_{\text{mean}} > 0, \\ -1 & \text{if } \mathbf{y}_{\text{mean}} < 0. \end{cases} \tag{4.2}$$

The final output of the algorithm can be summarized in the following results:

- the list of selected GO terms, ranked according to MCC values;
- the multiple classifier based on the classification models generated by the gene set annotated to the selected GO terms.

4.3 $\ell_1\ell_2$ regularization approach

Considering a binary classification problem on $n \times p$ ($p >> n$) gene expression data matrix X, for each observation i characterized by its gene expression profile $\mathbf{x}_i$=(x_{i1},...,x_{ip}), a linear classification method can predict the corresponding class label $\hat{y_i}$ using the following linear combination of the elements of $\mathbf{x}_i$:

$$\hat{y_i} = \beta_0 + \sum_{j=1}^{N} x_{ij}\beta_i = \bar{x}_i^T \cdot \bar{\beta} \tag{4.3}$$

This is the model of a single output $\hat{y_i}$, but in general $\mathbf{y}$ is a vector.
A model fitting procedure produces the vector $\hat{\beta}$=($\hat{\beta}_0$,...,$\hat{\beta}_p$) which represents the unknown weight coefficients assigned to each gene. A first simple but powerful approach for class prediction is the linear model fit by *ordinary least squares* (OLS), which estimates $\hat{\beta}$ by minimizing the residual sum of squares:

$$\hat{\beta} = \text{argmin}_\beta |\mathbf{y} - X\beta|_2^2 \tag{4.4}$$

However, the high number of features, which makes the system hugely underdetermined, and the redundancy of the feature set, which is common in high-dimensional problems, require methods able to deal with collinearities responsible for ill-conditioning. A classical way to solve these problems is provided by regularization methods. In these methods, feature selection and classifier construction are performed simultaneously by computing $\hat{\beta}$, estimate of β that minimizes a penalized objective function, allowing components of estimated β which are equal to zero. Feature selection is thus achieved, since only variables with nonzero coefficients will be used in the classifier.
Specifically, $\hat{\beta}$ is defined as:

$$\hat{\beta} = argmin_\beta \{m(\beta; D) + \gamma \times pen(\beta)\} \tag{4.5}$$

where D represents the dataset consisting of (x_1,y_1),...,(x_n,y_n) and m is referred to as the "classification objective function", which in the following is considered as the same used by the ordinary least squares approach ($|\mathbf{y} - X\beta|_2^2$).
The penalty pen(β) controls the complexity of the model. With the penalty function and properly chosen γ, some components of $\hat{\beta}$ are exactly zero. This leads to sparse classifiers and feature selection.
The tuning parameter γ >0 balances the goodness-of-fit and complexity of the model. When γ →0, the model has better goodness-of-fit. However, since the

classifier is too complex, it may have unsatisfactory prediction and be less interpretable. When $\gamma \to \infty$, the classifier has fewer input variables in it. The case of $\gamma{=}\infty$ corresponds to the simplest classifier where no input variable is used for classification.

There are different types of penalty functions. One of the best known strategies consists in adding a quadratic penalty, namely the ℓ_2 norm of the coefficient vector ($|\beta|_2^2 = \sum_j |\beta_j|^2$), to the loss function, *i.e.* to find the minimizer of the following penalized least squares objective function:

$$\phi_{ridge}(\beta) = \left[|\mathbf{y} - X\beta|_2^2 + \epsilon\,|\beta|_2^2\right] \tag{4.6}$$

where ϵ represent the regularization parameter γ. This estimate is known as the *ridge regression* [113, 114].

Since the objective function is strictly convex, it admits a unique minimizer given by

$$\beta_{ridge} = (X^T X + \epsilon I)^{-1} X^T \mathbf{y} \tag{4.7}$$

where I represents the $(p \times p)$ identity matrix.

Linear regularization methods select the relevant (stable) components of the solution and discard the others only on the basis of spectral properties of the matrix $\mathrm{X}^T\mathrm{X}$, independently of the data or response vector $\mathbf{y}$. The error level and therefore the output $\mathbf{y}$ is taken into account only according to the choice of the regularization parameter ϵ. However, since the eigenvectors are linear combinations of all the components of β, such methods are unable to perform variable selection, since all the weight coefficients are non-zero. This drawback can be overcome using thresholding or nonlinear shrinkage. A simple instance of a nonlinear regularization method is obtained when replacing in the ridge objective function by the ℓ_1-norm of the vector of the regression coefficients: $|\beta|_1 = \sum_{j=1}^{p} |\beta_j|$. This method is called *Lasso regression* (Least Absolute Shrinkage and Selection Operator) [115]. The minimization of the resulting objective function can be defined as:

$$\phi_{Lasso}(\beta) = \left[|\mathbf{y} - X\beta|_2^2 + \tau\,|\beta|_1\right] \tag{4.8}$$

with τ as regularization parameter. Such formulation increases the penalty on small coefficients (those for which $|\beta|_j$ <1) and decreases the penalty on large coefficients (those for which $|\beta|_j$ >1). In this way, this approach promotes solutions characterized by few large coefficients instead of many small ones. Moreover, the ℓ_1-penalty allows zero values instead of small ones, *i.e.* it favors *sparse* solutions. However, the Lasso is in general not variable selection consistent in the sense that small changes in the components of the input data X lead to a different feature

selection, typically with no appreciable change in the overall expected risk (or accuracy in the performance) of the obtained model. Thus, when the inputs are affected by noise or the number of examples is small compared to the number of features, the selection of the components of the model vector β might be driven by random fluctuations. Zou and Hastie [116] showed that when there are highly correlated input variables, such as in gene expression data, Lasso approach tends to select only one of the correlated variables. A penalty function able to effectively deal with high correlations is the *elastic net* penalty, which uses both a lasso-type and a ridge-type penalty in order to select groups of correlated variables. The minimization problem is then:

$$\phi_{en}(\beta) = \left[|\mathbf{y} - X\beta|_2^2 + \tau\,|\beta|_1 + \epsilon\,|\beta|_2^2\right] \tag{4.9}$$

With the pure ℓ_1 penalty, corresponding here to the case ϵ=0, in the case of correlated features the Lasso estimator can produce very different solutions, either by selecting one of them with a great weight or a few or else picking them all. The Lasso objective function is perfectly indifferent to the way the appropriate weight is distributed among those correlated features since all situations yielding the same value for the ℓ_1-norm are perfectly equivalent. On the other hand, the ℓ_2 penalty forces democracy in such cases since the ℓ_2-norm of the equal-weight configuration is smaller than that of all other configurations. However, the pure ℓ_2 penalty ($\tau = 0$) does not allow to perform variable selection The advantage of the combination of the two penalties is to both select variables and, among groups of correlated variables, to force democracy [15]. Empirical evidence indicates that the formulation described in Equation 4.9 produces stable solutions and exhibits an interesting grouping effect by selecting correlated features due to the presence of the ℓ_2-norm term, but suffers from the solution bias due to the shrinkage phenomenon induced by the ℓ_1-norm term). Moreover, good generalization performances are reported only for large values of the ϵ parameter, thus obtaining solutions very similar to those obtained by the ridge regression approach. In order to contrast bias and enhance the ability of ℓ_1-norm of promoting sparse solutions, De Mol *et al*. [112] proposed an approach which produces gene signatures able to effectively address prediction problems from high-throughput data like DNA microarray. The method learns from the available data a minimal set of genes whose expressions are best suited to accurately predict the biological parameter related to the problem at hand. By selecting the model through the combination of two optimization schemes, elastic net and regularized least squares. In particular, the first optimization procedure erforms gene selection by minimizing Equation 4.9

for a small value of the ℓ_2-norm parameter ϵ, whereas the second optimization is a regularized least squares (RLS) and the minimization problem is:

$$\phi_{RLS}(\beta) = \left[\left| \mathbf{y} - \tilde{X}\tilde{\beta} \right|_2^2 + \lambda \left| \tilde{\beta} \right|_2^2 \right] \tag{4.10}$$

where $\tilde{\beta}$ and $\tilde{X}$ represent respectively the weights vector β and the input matrix X, restricted to the genes selected by the first procedure. In this way, the method leads to a model which, unlike the elastic net alone, is characterized by both sparsity and low bias.

4.4 Classification schema and implementation

In the regularization step, the parameter ϵ (which controls the correlation among the features by the ℓ_2 norm) is a priori fixed equal to 10, allowing a high level of correlation among the selected features. The parameters for the feature selection τ (ℓ_1 norm) and λ, used in the optimization algorithm described in [112] to stabilize prediction results, are estimated on a varying geometric range of values opportunely chosen by the algorithm.
Different classification strategies were adopted in the learning process for each GO node. In a first preliminary work [117], $\ell_1\ell_2$ classification and feature selection were performed using a 5-fold cross validation strategy for parameter estimation, using an external test set for validation phase. In order to increase classification performance and to provide a more robust selection of biological functions, a bootstrap approach was also tested, using a resampling schema with B=100 external training/test splits with 5-fold cross-validation as internal resampling for parameter estimation. Final feature selection was performed by selecting those features which present non-zero weights in at least 60% of bootstrap samples. In order to provide a single classifier for each GO node, the parameter λ, used to calculate the weights β for the prediction of new subjects used in the regularized least squares step, was set as the median value of the its optimized estimations over the B samples.
The algorithm was implemented in Python language, using the library *l1l2py* (`http://pypi.python.org/pypi/L1L2Py`) for the classification method.

4.5 Results

In the following section, both results obtained in the preliminary analysis [117] and using the bootstrap approach are shown, in terms of both prediction accuracy

and interpretability through semantic similarity of selected GO terms. The two GO categories Biological Process (GOBP) and Molecular Function (GOMF) were used to define the gene sets.

4.5.1 Data

To better appreciate the various facets of the proposed approach, the method was applied on three real microarray datasets monitoring breast cancer patients with positive and negative estrogen receptor status were used (Table 4.1). Preprocessing steps such as background subtraction, probe cell normalization and expression level calculations, were performed using quantile normalization and Robust Microarray Analysis (RMA) software [84].

Datasets	Samples	ER+ Samples	ER- Samples
GSE2990 [80]	116	83	33
GSE3494 [81]	155	131	24
GSE7390 [82]	152	103	49

Table 4.1: **Breast cancer datasets used for the classification.**

4.5.2 Classification performance

The work described in [20] proved an improvement "on average", comparing classification performance between the standard approach of the $\ell_1\ell_2$ algorithm and the GO-based strategy. For each dataset, ten random splits of the data into training and test set were applied, with sample proportions 2/3 and 1/3 respectively. The ten splits were considered *independently*. In this case, feature selection for each node was directly performed by the $\ell_1\ell_2$ algorithm by the 5-fold cross-validation phase, since the classification is based on a single external split and the test set was used to calculate the MCC. Table 4.2 shows the average MCCs and the standard deviations obtained from the test sets of the ten splits in the three datasets. The MCCs of the ten splits obtained applying the new method on Biological Process and Molecular Function categories are significantly higher in both categories (p-value always lower than 0.021, Wilcoxon test) with respect to the MCCs obtained with the standard approach. Figure 4.1 displays also the distribution of MCC values over the ten replicates, highlighting the higher performance of GO-based methods with respect to the standard approach. Starting from these preliminary results, the bootstrap resampling schema described in

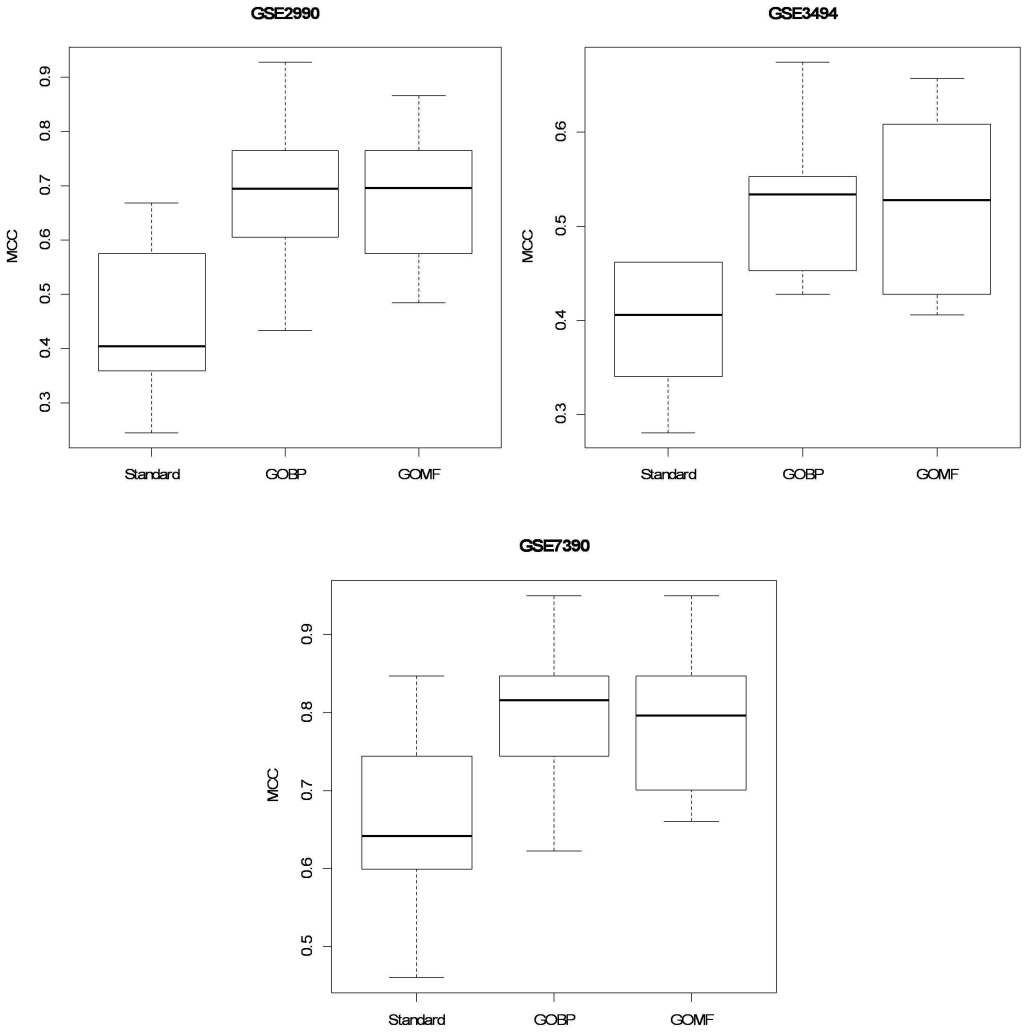

Figure 4.1: **Boxplot of MCC distribution in the three breast cancer datasets.** MCC values are displayed for the standard approach and the GO-based approach on GOBP and GOMF categories.

		Standard Method	GOBP based	GOMF based
GSE2990	Mean	0.445	0.685	0.679
	SD	0.137	0.152	0.121
GSE3494	Mean	0.396	0.523	0.525
	SD	0.069	0.076	0.088
GSE7390	Mean	0.661	0.798	0.796
	SD	0.117	0.104	0.094

Table 4.2: **Classification performance (MCC) over the ten random splits of the three breast cancer datasets.**

section 4.3 was performed. Since the MCC value obtained from the test set is used to decide if a GO term can be selected by the algorithm, the best strategy to assess classification results is to have an external validation test set which is independent from the selection of GO nodes. However, the limited number of samples in the ER- class in each dataset do not allow an independent blind test. The three datasets represent three independent studies of the same clinical classification problem. Thus, a between dataset assessment was performed with the new analysis. Table 4.3 shows the MCC values obtained on the six combination of the three breast cancer dataset. On average, GO-based approach outperforms the standard method when subjects from a different study (thus carried out in a different laboratory with a different experimental protocol) are used.

Training	Test	Standard	GOBP based	GOMF based
GSE2990	GSE3494	-0.05	0.49	0.52
GSE2990	GSE7390	0.23	0.76	0.37
GSE3494	GSE2990	0.09	0.5	0.41
GSE3494	GSE7390	0.29	0.3	0.59
GSE7390	GSE2990	0.31	0.42	0.4
GSE7390	GSE3494	0.2	0.17	0.23

Table 4.3: **Between-dataset classification performance (MCC).**

4.5.3 Interpretability of GO lists

Beside classification performance, it is also interesting to assess the interpretability of the obtained results. The work described in [117] pointed out the stability of results obtained by the GO-based approach considering those GO

terms with a frequency equal or higher than 0.8 in the ten splits. Many of the selected GO terms were common to all the three breast cancer datasets, in particular the GO terms related to "response to oxidative stress", "developmental process" and "regulation of cell proliferation" in Biological Process and the GO terms related to "oxidoreductase activity" and "metal ion binding" in Molecular Function. However, it is useful to compare how the GO lists selected by the proposed method are stable across the three datasets with respect to a simple enrichment approach. Three different analysis pipelines were compared: the univariate method SAM with GO enrichment, the standard application of $\ell_1\ell_2$ classification method with GO enrichment and the new approach with the bootstrap approach.

		GSE2990 vs GSE3494	**GSE2990 vs GSE7390**	**GSE3494 vs GSE7390**
SAM	GOBP	0.511	0.435	0.452
	GOMF	0.306	0.439	0.392
Standard	GOBP	0.460	0.687	0.531
	GOMF	0.316	0.576	0.331
GO-based	GOBP	0.512	0.690	0.786
	GOMF	0.653	0.783	0.654

Table 4.4: **Semantic similarity levels obtained by between dataset analysis.**

In order to account for information redundancy as in the proposed method, the elim approach proposed in [111] was chosen as enrichment method; the selection of enriched GO terms was performed by applying on resulting p-values the same threshold used by the elim approach, which is fixed to 0.01 by default. For each analysis pipeline, the three lists of selected GO terms obtained for the three breast cancer datasets were compared using the semantic similarities described in Chapter 5 (section 5.2.5), which allow to quantitatively assess the biological coherence of selected biological processes and molecular functions. In particular, given two lists of GO terms, the best-match average approach is used as a measure of similarity between the biological functions of the two GO lists (Equation 5.10). This method provides a score ranging between 0 and 1, where 1 means that the two GO lists have exactly the same GO nodes.
Table 4.4 shows that, even if standard $\ell_1\ell_2$ approach improves performance obtained by the SAM approach in two cases, GO-based approach is the best per-

former, providing similarities ranging between 0.51 and 0.79 in the three pairwise comparisons.

4.6 Discussion

When comparing classes of subjects belonging to different phenotypes using high-throughput technologies such as microarrays, biological annotation of differential features can give immediate and intuitive information on the phenomenon under investigation. However, organizing results in a structured way is not straightforward. A new method was here presented to integrate biological annotation into statistical class prediction analysis of microarray data. Differently with respect to the previous studies proposed in the literature, the method exploits the GO graph on a global scale taking into account the redundancy of information characterizing GO annotations and generates a multiple classifier built on the most predictive biological pathways selected by the algorithm. Compared to standard classification analysis, the method significantly improves prediction accuracy. The improvements highlighted by the presented results are probably due to the limited number of genes, restricted to those belonging to a single GO term, used to build each classifier: in this way, the curse of dimensionality effect is reduced and a more robust statistical analysis is promoted. Moreover, the method is able to organize results into subsets of genes both 1) highly correlated (from the $\ell_1\ell_2$ classification approach) and 2) annotated to groups of GO terms with similar meaning. The list of GO terms provided as output gives a functional-based characterization of the disease in an easy-to-read way, selecting more stable biological functions and allowing a better interpretability of results when different datasets are analyzed. Instead of GO, it would also be possible to use different functional groups, such as KEGG pathways or, for example, or to annotate genes in different group depending on which are their known transcription factors. Here, Gene Ontology was chosen because it allows annotating the largest number of genes in comparison with other functional annotation criteria and it represents the best hierarchical structured biological information currently available. However, different information can be organized in a hierarchical way and exploited by the proposed method to manage redundancy of information.

Chapter 5

Improving biomarker list stability by integration of biological knowledge in the learning process

5.1 Background

As observed in the previous chapters, there are two stability issues arising in gene expression classification and analysis. First, since training data are often scarce, predictive models obtained from different datasets can be extremely different. Secondly, since the number of features is generally very high, then features can be combined in many different ways to give solutions able to explain the data, allowing many possible sets of features equally good in terms of the accuracy.
The bootstrap approach, extensively assessed in the second chapter, has been demonstrated helpful in addressing the first issue. However, this method does not solve the problem of the instability due to the high number of features. In fact, the crucial problem is that the classification task is under constrained. To address this issue, additional information available on the relationships between genes should be used to improve biomarkers lists stability, taking into account the complex gene relationships, instead of considering genes as independent features. In the previous chapter, Gene Ontology (GO) structure has been explored as possible prior information to be integrated in the learning process. Besides functional annotations, other different types of biological background knowledge exist, depending on the level of the system that is described. Some databases focus only on the biological interactions among proteins, that define the processes within a living cell and summarize these in usually manually curated pathway models,

e.g. Human Protein Reference Database (HPRD) [47]. Another useful type of information can be extracted from different studies stored in repositories like Gene Expression Omnbibus (GEO) [118] or ArrayExpress [119]. Correlations characterizing possible interactions among genes can be obtained by meta-analysis on a set of gene expression datasets. All these types of information were explored recently by different studies proposing new approaches for incorporating external biological knowledge into risk prediction models and several efforts in this direction were recently presented in the literature. In [29], pathway information was incorporated into the biomarker discovery process using available protein-protein interaction networks and considering subnetworks as features. Logistic regression models were applied on expression profiles of two cohorts of breast cancer patients and results were assessed in terms of both agreement between subnetworks identified in the two datasets and classification accuracy. Rapaport *et al.* [120] proposed a formulation of support vector machines (SVM) to estimate a predictive model by constraining the weights of connected genes to be similar, allowing to associate positive or negative contributions to regions of the network. In [121, 122, 123] other topological properties of KEGG pathways or networks reconstructed from gene expression data were used to constrain the learning process. In particular, [123] used regularization and integrated prior knowledge defining KEGG pathway based penalty terms. All the above methods focused on prediction performance, without considering in a systematic way the stability issue. Recent works started considering the problem of biomarker list stability [64], but an overview of the ability of different sources of biological knowledge to improve the reproducibility of biomarkers lists is not yet available in the literature.

In this chapter a new method able to integrate prior knowledge in the learning process is presented and, differently from previous works, the performance obtained by different sources of prior knowledge are compared in terms of biomarker lists stability. In particular, the method proposes a standardized way to incorporate in a kernel function different types of biological knowledge like functional annotations, protein-protein interactions, and expression correlation among genes, with the only constraint that the information is codified by a similarity matrix. The feature space is then transformed such that the more similar two features are, the more closely they are mapped. Other studies describe different approaches that integrate different datasets [124] by combining kernels [38] to improve classification performance and robustness of the results. Here, a different and maybe complementary aspect of the problem is handled by incorporating different types of knowledge, allowing the proposed method to be used with any kernel method.

5.2 Integration of prior knowledge in the learning process

Expression data are given as very high dimensional vectors of measurements. The high dimensionality makes the task of biomarkers discovering very hard. This is especially due to the fact that the task is under constrained. A linear transformation of the examples (*i.e.* the biological samples) is performed in a way that classifiers computed on transformed examples have a higher stability, hopefully preserving the accuracy. This transformation is made by using prior biological information about genes in a way to maintain the structure of the problem. In the following section, a linear classifier resembling the *Bayes Point Machine* [125] and the embedding of knowledge data into feature spaces are introduced. Then, the algorithm implementing the proposed approach is described.
In the following, the examples are denoted by $\{\vec{x}_1, ..., \vec{x}_M\}$, *i.e.* the N dimensional vectors of expression data obtained for M subjects, where N is the number of genes. Each example has associated a binary label y_m (m=1,...,M) having values in -1,+1.

5.2.1 Linear classifier

Given a linearly separable classification task, there are in general infinitely many linear classifiers (hyperplanes) that can correctly classify the examples. This set is commonly called the *version space*. When the number of features is very high, the version space tends to have a large volume. Formally, the version space for linear classifiers can be defined as:

$$V = \{\vec{w} | y_m(\vec{w} \cdot \vec{x_m}) > 0, for\ all\ m = 1, ..., M, \|\vec{w}\| = 1\} \tag{5.1}$$

Without any loss of generality, weights are considered with unitary norm. A very popular algorithm to find a linear classifier which correctly separates the training examples (*i.e.* an element of the version space) is the Perceptron algorithm [36] which can be briefly described as in the following. Assuming the training vectors x and w of size N, with w initially set to the zero vector, the algorithm runs in *epochs*. On each epoch all the training examples x_i, for i=1,..,M, are used in the algorithm and the vector w is updated whenever the associated classifier makes a mistake on x_i, *i.e.* if $(y_i sign(w \cdot x_i) \leq 0)$ then $w = w + y_i x_i$. When the training set is linearly separable, the perceptron is guaranteed to eventually converge to a vector (hyperplane) which correctly separates the training data, *i.e.* the solution

is an element of the version space.
It can be shown that other kernel based algorithms, like for example the hard version of SVM (Section 2.3.2), also have solutions in the version space. In the particular case of SVM this solution is in fact unique and is the one which maximizes the margin on the training set [34]. As shown in [125], the center of mass of the version space, the so called *Bayes point* (Bp), would be the optimal choice, even better than SVM (which can be considered an approximation of the Bp), with nice theoretical properties in terms of its generalization ability. An algorithm that approximates this optimal Bp solution is the so called *Bayes point machine*, which considers the average of the solutions of several runs of the perceptron. A variant of this algorithm was considered in the analyses.
When a feature space is characterized by high dimensionality and the features are considered independent, *i.e.* there are many more variables (features) than constraints (examples), the task is under constrained. This often implies that the version space volume is large and can change extremely both in form and size, depending on which examples are used for training. It is clear that this produces instability. Section 5.2.3 will explain how to add available domain knowledge to introduce structural constraints in the problem in order to improve robustness of a linear classifier.

5.2.2 Feature ranking

The values w_i of a linear classifier represent the degree of importance and the bias that a given feature i provides to the decision. High positive (negative) values tell us that such feature is important to classify an instance as positive (negative). For this reason, the absolute value of the weights can also be used as a criterion for feature ranking.

5.2.3 Similarity matrix integration

When prior knowledge is available providing information about gene-gene similarity, this knowledge can be effectively used by mapping examples into a feature space where linear solutions preserve these similarities.
Considering a linear transformation of the data via a matrix P, *i.e.* $\phi(x) = Px$, do exist desirable properties of ϕ which make the task of discriminating positive versus negative examples simple enough in the target space? It is well known that a measure of the goodness of an embedding is the ratio between the maximal norm R, the highest norm (or length) of any example x_m, and the margin γ

of the examples, namely $G = (R/\gamma)^2$.
For separable data, the margin is defined as the distance between the optimal separating hyperplane and the examples. In the case of perceptron classifiers, the value G is also related to the number of mistakes the perceptron algorithm makes to converge [36]. These considerations seem to indicate that the margin of transformed examples should be large in order to get high performance. However, when the expected margin (or equivalently, the expected volume of the version space) is too large, it generally leads to unstable solutions for small datasets. A solution, which represents a trade-off between these two (apparently) opposite goals, is to choose an embedding of data where norm of vectors are as small as possible but data remain linearly separable.
Specifically, the proposed strategy makes a linear embedding of data via a bi-stochastic matrix. Here, stochastic matrices are considered because they have the property to map vectors x into shorter ones (compression) and thus to make the maximal norm R of target examples smaller (this is due to the fact that the eigenvalues of a stochastic matrix are all in [0,1]). As previously seen, this together with large margin solutions guarantee a good performance of the embedding.
Let S be a symmetric similarity matrix with elements in [0,1] with 1's in the diagonal, the associated stochastic matrix P is obtained as in the following:

$$P = D^{-1}(I + \alpha(S - I)) \tag{5.2}$$

where I is the identity matrix, D is a diagonal matrix with elements corresponding to sums of elements in the rows/columns of $(I+\alpha(S-I))$, and α¿0 is a tuning parameter. Note that when α=0, then P=I and the feature space coincides with the original space. The parameter α is fixed according to the best stability performance, measured by the Canberra distance (Equation 2.15).
Given a perceptron-like solution in the target space, the weight vector can be expressed as a weighted sum of the examples in feature space, namely $\vec{w} = \sum \beta_m \phi(\vec{x}_m)$, and the following holds:

$$\begin{aligned} |w_i - w_j| &= \left| \sum_m \beta_m \sum_k P_{ik} xmk - \sum_m \beta_m \sum_k P_{jk} x_{mk} \right| = \\ &= \left| \sum_m \beta_m \sum_k (P_{ik} - P_{jk}) x_{mk} \right| = \\ &= \left| \sum_k (P_{ik} - P_{jk}) \sum_m \beta_m x_{mk} \right| = \\ &= \left| (\vec{P}_i - \vec{P}_j) \cdot h \right| \leq c \left\| \vec{P}_i - \vec{P}_j \right\| \end{aligned} \tag{5.3}$$

where $c \leq 0$ is a constant which does not depend on indices i and j. Thus, the matrix P can be seen as a coding matrix for genes. Specifically, the i-th gene is codified by P_i. This result shows that when two genes have similar codes, the difference in the weight vector cannot be too large. It is important to note that this result does not imply that the same gene will have the same position in the ranking generated by two independent experiments, *i.e.* that the same biomarkers will be selected. The result above simply means that the relative position of two similar genes will be similar in the two experiments. However, if the matrix P contains reliable information, this should hopefully produce similar lists of biomarkers.

5.2.4 Classification algorithm and biomarker list generation

The proposed algorithm is based on the perceptron algorithm and resembles the Bayes point machine. The implementation of the classification algorithm is available at the link: `http://www.math.unipd.it/~dasan/biomarkers.html`. The algorithm starts by mapping data using the matrix P. The transformed data are standardized by subtracting from each gene expression value its mean across the samples and dividing by its standard deviation. Then, data are randomly split (70% training, 30% test) for a number T=1000 of times. For each one of these splits a run of the perceptron algorithm is performed on its training data (to increase randomization data are also shuffled before each perceptron epoch). Thus, for each split t, a weight vector wt is obtained and normalized to unitary norm. For each split, the accuracy a_t is also evaluated with respect to the test partition. The final solution is obtained as the average of weight vectors w_t, *i.e.* $W = AVE(w_t)$.

Note that the expected accuracy of W on new unseen examples can also be estimated by using available data with the following method. Let Q be the design matrix with entries Q_{tm}=1 if the example x_m is in the training partition of split t, and 0 otherwise. For each example x_m a predictor $W(m) = AVE(w_t)$ is built using just the weights w_t such that Q_{tm}=0, *i.e.* taking the average of the weight vectors for the construction of which the example x_m was not used. Finally, the classifier $W(m)$ is tested against x_m.

The accuracy observed by applying this method on all available data is an estimate of the expected accuracy of W. The list of biomarkers returned by the algorithm is the list of genes ordered according to the absolute value of their correspondent value in W.

The method described above can also be seen as a leave-one-out estimate of the accuracy. However, the same method can be easily adapted to a (k-fold) cross-validation type of analysis. In this case, the overall procedure would change as in the following:

- Split data in k sets X_1,..,X_k;
- Train models W_1,..,W_m, where W_t, t=1,..,k, is learned on the set $X \setminus X_t$, with the method presented above, and get the accuracy $ACC(X_t)$ on the set X_t;
- Evaluate the overall accuracy as the average of these partial accuracy estimates.

The advantage of using a k-fold type of analysis instead of the leave-one-out type of analysis is its lower variance for small samples. The disadvantage is that the method is more computational demanding. Some experiments were done using both methods and no significant differences were observed in the obtained results.

5.2.5 Similarity matrices

Three different kinds of data were considered as prior knowledge to be integrated in the feature ranking: 1) Gene Ontology functional annotations; 2) the network of protein-protein interactions; 3) gene expression profiles from a collection of breast cancer studies. All these data were used to calculate different kinds of similarity measures s_{ij} between pairs of features i and j based on:

- Semantic similarity of functional annotations;
- Topological similarity in the network of protein-protein interactions;
- Correlation between gene expression profiles.

The corresponding similarity matrix S for N variables is the symmetric $N \cdot N$ matrix whose element s_{ij} refers to the similarity between the features i and j. In the following, the methods for codifying the three types of prior knowledge into the corresponding similarity matrices are described in details. Since Affymetrix data were considered, indexes i and j refer to probesets. What follows can be easily generalized to consider genes or proteins. Each subsection first describes the biological information and then illustrates the metrics used to generate the corresponding similarity matrix.

Semantic similarity on GO annotations

Semantic similarity measures are used to evaluate the degree of relatedness between two features by assigning a metric based on the likeness of the semantic content of their GO annotation, defined over the GO terms associated with the genes. One early idea was to define the similarity as a function of the distance between the two terms in the ontology graph [126] or the length of their common path from the root, *i.e.* the number of common parents [127]. However, pure graph-based similarities suffer from the fact that the depth of a term within the ontology is not necessarily indicative of its specificity [128]. This motivated the formalization of the notion of specificity with the definition of the Information Content (IC) of a given term:

$$IC(t) = -log\frac{f(t)}{f(root)} \tag{5.4}$$

where $f(t)$ is the number of occurences of the annotations associated to the term c and its descendants, estimated as:

$$f(t) = |annot(t)| + \sum_{c \in children(t)} f(c) \tag{5.5}$$

The root term, which is implied by all terms, has an IC equal to 0. By contrast, rare terms have a high IC. Thus, the intuition behind the use of the IC is that the more probable a concept is, the less information it conveys.

Resnik [129] combined the notion of IC with the ontology structure to define the similarity the similarity between two concepts t and u as the information content of the most informative common ancestors, $MICA(t, u)$. Formally:

$$Sim_{Resnik}(t, u) = max_{c \in MICA(t,u)} IC(c) \tag{5.6}$$

The more informative is the common ancestor, the greater the information shared by the concepts, and consequently their similarity. An inconvenient aspect of this measure is that it is not normalized. To overcome this, Lin normalized the measure between 0 and 1 [130]:

$$Sim_{Lin}(t, u) = \frac{2 \times Sim_{Resnik}(t, u)}{IC(t) + IC(u)} \tag{5.7}$$

A warning related to the use of this type of measure is that, as an effect of the normalization, genes annotated to general terms tend to have higher similarities on average than genes annotated to specific terms. This phenomenon, referred to as the "shallow annotation problem", has been discussed in the literature [131].

Even though it may be affected by this problem, Lin's formula has already been proven to outperform other algorithms [132].
So far, semantic similarity measures for pairs of GO terms have been presented. However, the analysis is at the level of genes or their products, each associated with one or more terms. A first approach to calculate the semantic similarity score between two genes g_1 and g_2 annotated in the database consists in considering all possible pairs of GO terms associated with both genes and to use either the maximum or the average similarity as measures for the two gene:

$$Sim_{max}(g_1, g_2) = max_{t \in GO_1, u \in GO_2} Sim(t, u) \tag{5.8}$$

or

$$Sim_{avg}(g_1, g_2) = avg_{t \in GO_1, u \in GO_2} Sim(t, u) \tag{5.9}$$

where GO_1 and GO_2 are the groups of GO terms associated to genes g_1 and g_2, respectively.
However, both variants have flaws. The maximum approach can answer the question of whether two gene products share a functional aspect, but is unsuitable to assess their global similarity: it is indifferent to the number of functional aspects they share and to the number of functional aspects in which they differ, since genes that differ in all but one functional aspect will still show a high similarity under this measure. On the other hand, the average approach makes an all-against-all comparison of the terms of two gene products and produces counterintuitive results for gene products that have several distinct functional aspects, because the average tend to be dominated by pairs of different GO terms: considering a gene with n GO terms, the number of pairs involving identical GO terms scales linearly in n, but the number of pairs involving different terms scales quadratically [133].
A good balance between the maximum and the average approach is the best-match average (BMA) [134], which computes the average over the reciprocal best matching pairs only. This measure was chosen in the analysis to build similarity semantic matrices representing Gene Ontology information. Thus, using this approach in combination with Lin's similarity measures between the GO terms, the semantic similarity scores s_{ij} between two features i and j are calculated as:

$$s_{ij} = \frac{\frac{1}{GO_i} \sum_{t \in GO_i} max_{u \in GO_j} Sim_{Lin}(t, u) + \frac{1}{GO_j} \sum_{u \in GO_j} max_{t \in GO_i} Sim_{Lin}(u, t)}{2} \tag{5.10}$$

where GO_i and GO_j are the groups of GO terms t and u associated to the features (*e.g.* probesets, genes) i and j, respectively.

Here, Molecular Function and Biological Process GO annotations related to the probesets were downloaded from NetAffx database (`http://www.affymetrix.com/analysis/index.affx`), while the DAG structure was extracted from the Bioconductor package GO.db.

Topological similarity on protein-protein interactions

Many recent studies have demonstrated the topology of network offers effective information for a better understanding of gene products molecular functions [135, 136] and the underlying mechanisms describing how they interact. Different useful topological metrics were proposed to measure similarity of a protein pair in protein-protein interactions (PPIs).

Before describing the similarity measures which are most used in the literature, some terms and notations are first introduced. The network of the interactions is defined as graph G=(V,E) consisting of a set of nodes V and a set of edges E between them; p_i and p_j refer to proteins which are the nodes of the network, whereas $N(p_i)$ and $N(p_j)$ are the neighbors of p_i and p_j respectively, and $N(p_i, p_j) = (N(p_i) \cap N(p_j))$.

Normalized geodesic distance

The normalized geodesic distance (NG) between two proteins p_i and p_j is defined as the normalized length of the shortest path, $l(path(p_i, p_j))$, from p_i and p_j, obtained by dividing $l(path(p_i, p_j))$ by the maximum of the shortest paths between all pairs of proteins. The similarity $s(p_i, p_j)$ between two proteins is derived as 1 minus the normalized shortest path:

$$s(p_i, p_j) = 1 - \frac{l(path(p_i, p_j))}{max_{p_k, p_r \in V(G)} [l(path(p_i, p_j))]} \tag{5.11}$$

Jaccard coefficient

Since proteins, which share more common neighbors are likely to share similar biological characteristics, neighbors counting method has been mostly studied and widely used in protein function prediction. One of this is the Jaccard coefficient (JA) [137], which is defined as the ratio between the number of neighbors which two proteins share (common neighbors) and the total number of proteins they are connected to:

$$s(p_i, p_j) = \frac{|N(p_i, p_j)|}{|N(p_i) \cup N(p_j))|} \tag{5.12}$$

Functional similarity

Chua *et al.* [138] proposed a measure named Functional Similarity (FS) for measuring the common neighborhood similarity of two proteins p_i and p_j in an interaction network G. For an un-weighted network (0/1 weights), this measure can be defined as:

$$s(p_i, p_j) = \frac{2|N(p_i, p_j)|}{|N(p_i) - N(p_j)| + 2\,|N(p_i, p_j)| + \lambda_{ij}} \times \frac{2|N(p_i, p_j)|}{|N(p_j) - N(p_i)| + 2\,|N(p_i, p_j)| + \lambda_{ji}} \quad (5.13)$$

where

$$\lambda_{ij} = max(0, n_{avg} - (|N(p_i) - N(p_j)| + 2\,|N(p_i, p_j)|)) \quad (5.14)$$

n_{avg} is the average number of neighbors of each protein in the network. The purpose of the λ_{ij} factor is to penalize the score between proteins pairs where at least one of the proteins has too few neighbors, since the score may not be very reliable in such a case.

Essentially, FS separates the functional similarity of two proteins into two probabilities that denote the conditional probabilities of p_i and p_j being functionally related given the neighborhoods of $N(p_i)$ and $N(p_j)$, respectively. Each of these conditional probabilities are computed as how similar the set of common neighbors of p_i and p_j ($N(p_i, p_j)$) is to the two sets of individual neighbors $N(p_i)$ and $N(p_j)$. The final FS score is obtained as a product of these probabilities, assuming that they are independent. For the computation, FS assumes that a protein, i.e p, is included in its direct neighborhood, *i.e.* $N(p)$.

Probabilistic common neighborhood similarity

A probabilistic measure for the statistical significance (SC) of the common neighborhood configuration of two proteins p_i and p_j has been recently proposed by [139]. The measure is defined as the negative logarithm of the probability of p_i and p_j having a certain number of common neighbors by random chance:

$$s(p_i, p_j) = -log_{10}(prob(N, |N(p_i)|, |N(p_j)|, |N(p_i, p_j)|)) \quad (5.15)$$

Here, N is the total number of proteins in the network, and $prob(N,|N(p_i)|,|N(p_j)|, |N(p_i, p_j)|)$ is computed on the basis of the Hypergeometric distribution:

$$prob(N, |N(p_i)|, |N(p_j)|, |N(p_i, p_j)|) = \sum_{k=|N(p_i,p_j)|}^{min(|N(p_i)|,|N(p_j)|)} \frac{\left(\frac{|N(p_i)|}{k}\right)\left(\frac{|N|-|N(p_i)|}{|N(p_j)|-k}\right)}{\left(\frac{|N|}{|N(p_j)|}\right)} \quad (5.16)$$

Thus, the higher the probability (5.16), the higher the value of $s(p_i, p_j)$ is.
Equations (5.11), (5.12), (5.13) and (5.15) are finally used to derive s_{ij}. Since different proteins can be associated to different probesets, the value of the similarity score s_{ij} between probesets i or j was obtained by averaging the similarity scores of the associated proteins:

$$s_{ij} = \frac{\frac{1}{P_i}\sum_{p_i \in P_i} max_{p_j \in P_j}\left[s(p_i, p_j)\right] + \frac{1}{P_j}\sum_{p_j \in P_j} max_{p_i \in P_i}\left[s(p_j, p_i)\right]}{2} \tag{5.17}$$

where P_i and P_j are the sets of proteins p_i and p_j annotated to the probesets i and j, respectively.
Topological information on PPI was extracted from HPRD [47], which contains manually curated scientific information pertaining to the biology of most human proteins and is the database that includes most human protein-protein interactions, as shown in [102]. The 22207 features in the considered datasets were mapped into 9521 proteins using RefSeq identifiers; this resulted in 37080 interactions.

Correlation based similarity

Gene expression profiles over the ten publicly available breast cancer microarray studies (Table 5.1) were compared using similarity measures based on Pearson correlation coefficient, Spearman rank correlation coefficient and Mutual Information, which provide a general measure to analyze dependencies in gene expression data [140, 141, 142].

Datasets	**Platform**	**Samples**
GSE2034 [143]	HGU133A	286
GSE6532 [144]	HGU133A / HGU133plus2	225
GSE11121 [145]	HGU133A	200
GSE2990 [80]	HGU133A	189
GSE1456 [146]	HGU133A	159
GSE7390 [82]	HGU133A	155
GSE5460 [147]	HGU133plus2	127
GSE3494 [81]	HGU133A	110
GSE5847 [148]	HGU133A	95
GSE4922 [149]	HGU133A	40

Table 5.1: **Breast cancer data sets used for the co-expression matrix.**

Pearson correlation coefficient

The Pearson correlation coefficient (PE) is the most widely used measurement of association between two vectors. Let x and y be the expression profiles of two probe sets which are analyzed in terms of degree of association. For pairs of quantities (x_i,y_i), $i = 1, \ldots, m$ the correlation coefficient ρ_{xy} is given by the formula:

$$\rho_{xy} = \frac{\sum_{i=1}^{m}(x_i - \mu_x)(y_i - \mu_y)}{(m-1)\sigma_x\sigma_y} \tag{5.18}$$

where μ_x, σ_x and μ_y, σ_y are sample means and standard deviations of x and y over the m measurements. The Pearson correlation reflects the degree of linear relationship between two profiles.

Spearman rank correlation coefficient

The Spearman rank correlation coefficient (SP) is a non-parametric measure of association that summarizes nonlinear relationship between two numerical variables. The Spearman correlation coefficient is computed as:

$$\rho_{xy} = \frac{\sum_{i=1}^{m}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{m}(X_i - \mu_X)^2 \sum_{i=1}^{m}(Y_i - \mu_Y)^2}} \tag{5.19}$$

where X_i, Y_i denote the ranks of x_i, y_i, respectively and μ_X,μ_Y the corresponding means. When there are no ties, the formula reduces to:

$$r = 1 - \frac{6\sum_{i=1}^{m} d_i^2}{n(n^2-1)} \tag{5.20}$$

where d_i is the difference between the values of X_i and Y_i.

Using both the Pearson and the Spearman correlation, the similarity s_{ij} between the expression profiles x and y of two probesets i and j was defined as:

$$s_{ij} = |\rho_{xy}| \tag{5.21}$$

Mutual Information

The Mutual Information (MI) provides a general measure for dependencies in the data. It is a well known measure in information theory and it has been widely used to analyze gene-expression data [140, 141, 142]. Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$I(X,Y) = \sum_{y \in Y}\sum_{x \in X} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{5.22}$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y, respectively.

To calculate MI, data were needed to be quantized on L intervals. There is no optimal solution to choose L, since it depends on data normalization and on the particular biological application [150]. As suggested in [151], heuristic lower/upper bounds on the number of intervals were considered [152, 151]: $MI_{low}=\lfloor 1 + log_2 m \rfloor$ and $MI_{up}=\sqrt{m}$, where m is the number of expression values. L was set equal to 25. The score s_{ij} was set to the value $I(X,Y)$ obtained from the quantized expression profiles X and Y of two probesets i and j.
Breast cancer datasets reported in Table 5.1 were extracted from GEO, selecting those with a medium to large sample size. Redundant subjects were removed. All datasets were hybridized using Affymetrix U133 Genechips^{TM} (HG-U133A and HGU133plus2) and were analyzed using A-MADMAN, an open source web application, which allows the retrieval, annotation, organization and meta-analysis of gene expression [153]. In particular, the software enables the integrative analysis of data obtained from different Affymetrix platforms through meta-normalization. Affymetrix chip definition files were used to annotate the arrays and gene expression intensity signal was normalized using RMA algorithm. The resulting gene expression matrix collects the expression levels of 21921 probesets over 1586 biological samples.

5.3 Data and evaluation of the biomarker lists

The three real microarray datasets monitoring breast cancer patients with positive and negative estrogen receptor status, described in the previous chapter (Table 4.1), were chosen for the analyses. Results were evaluated in terms of both stability of the biomarker lists obtained by the Canberra distance [7] and the accuracy performed by the perceptron classifier.
The Canberra distance, described in section 2.6, is a weighted version of the Spearman's footrule which considers the variations in lower portions of the lists less relevant than those in the top. Its normalized version can be obtained by dividing the distance in Equation 2.15 by its expected (average) value, approximated by $(log(4) - 1)p + log(4) - 2)$ for the complete lists. The normalized Canberra distance ranges between 0 (maximal stability) and 1.4 (maximal instability), with 1 in the case of randomly generated lists.
The average number of iterations needed by the perceptron in the algorithm is also considered as a good indicator of the ratio between the maximal norm of transformed vectors and the margin one can obtain in feature space. This value is considered as a measure of how much difficult is the transformed task.

Ranked features lists obtained using different similarity matrices were evaluated both within datasets, *i.e.* comparing the 1000 different lists obtained using bootstrap, and between datasets, *i.e.* comparing the global lists obtained by analyzing datasets GSE2990, GSE3494 and GSE7390. For the within dataset comparison, the Canberra distance was applied on the 1000 complete lists resulting from the bootstrap resampling schema adopted by the classification algorithm. For the between dataset comparison, the Canberra distance was applied on the sublists of length k, with k corresponding to the minimum Canberra distance within datasets (average of the three values obtained for the three datasets). Finally, for the best performing similarity matrices, the union of the sublists of length k obtained using the three datasets, where k ranges from 1 to the maximum number of features (22207), was considered in order to quantify the possible lack of consistency of the global lists.

5.4 Results

5.4.1 Within dataset assessment

Table 5.2 reports the average normalized Canberra distance and classification accuracy for all the three breast cancer datasets and for all similarity matrices. Results are reported for the cases where prior information is not used (α=0) and when for each similarity matrix the value of α (Equation 5.2) which minimizes the Canberra distance is used. For all the three datasets, all types of biological information are able to decrease the average normalized Canberra distance over the biomarker lists with respect to the standard classification approach. In particular, three types of prior knowledge are best performers in this task: Gene Ontology Biological Process (GO BP), Gene Ontology Molecular Function (GO MF) and protein-protein interactions codified by the normalized geodesic distance (PPI NG). For these three types of biological knowledge, the improvement in list stability, which ranges between 26% and 37%, is achieved without a corresponding loss in accuracy since this latter changes in a range between minus 2% to plus 3%.

Table 5.2 also reports the number of iterations needed by the classification algorithm to reach convergence, averaged across the 1000 bootstrap splits. Compared to other types of prior knowledge, the higher number of iterations are observed with the correlation (PE and SP) and Mutual Information (MI) based matrices, whereas PPI measures lead the classifier to reach convergence with a lower

	GSE2990	GSE3494	GSE7390
No prior	0.89 (95%)	0.93 (93%)	0.90 (98%)
	7	10	6
GO BP	0.62 (93%)	0.63 (95%)	0.60 (96%)
	15	21	13
GO MF	0.63 (93%)	0.68 (94%)	0.60 (97%)
	17	24	15
PPI NG	0.57 (94%)	0.58 (96%)	0.53 (97%)
	10	14	9
PPI JA	0.87 (95%)	0.91 (93%)	0.87 (97%)
	7	11	7
PPI FS	0.88 (95%)	0.92 (95%)	0.88 (97%)
	7	11	7
PPI SC	0.83 (95%)	0.86 (95%)	0.83 (96%)
	8	13	8
PE	0.78 (95%)	0.89 (96%)	0.79 (96%)
	49	56	37
SP	0.78 (95%)	0.89 (95%)	0.79 (95%)
	48	56	38
MI	0.76 (91%)	0.80 (94%)	0.73 (94%)
	130	207	131

Table 5.2: **Classification performance within breast cancer datasets.** Normalized Canberra distance between feature lists obtained for datasets GSE2990, GSE3494 and GSE7390, using the standard classification approach without prior knowledge integration and different prior knowledge based similarity matrices: Gene Ontology Biological Process (GO BP), Gene Ontology Molecular Function (GO MF), protein-protein interactions codified by the normalized geodesic distance (PPI NG), the Jaccard coefficient (PPI JA), the functional similarity (PPI FS), the probabilistic common neighborhood similarity (PPI SC), the Pearson correlation (PE), the Spearman rank correlation (SP) and the Mutual Information (MI). Predictive accuracy is indicated in brackets, whereas the number of iterations obtained by the classifier is reported below the other scores.

number of iterations, *i.e.* they improve class separability. However, except the normalized geodesic distance, all the other protein-protein interaction measures show the lowest gain in reproducibility.

5.4.2 Between dataset assessment

Table 5.3 reports the average Canberra distance obtained by comparing datasets GSE2990 vs GSE3494, GSE2990 vs GSE7390, GSE3494 vs GSE7390, and the resulting average Canberra distance together with the average classification accuracy across the three datasets for k corresponding to the minimum Canberra distance within datasets (average of the three values obtained for the three datasets). GO BP, GO MF and PPI NG are confirmed as the best performing kinds of prior knowledge. In addition, MI based similarity matrix shows performance comparable to the former similarity matrices. In order to better assess the improvement highlighted in these four similarity matrices, the size of the union sets of the biomarker lists of length k over all the three datasets is considered (Figure 5.1). The more two lists are similar, *i.e.* containing the same features, the more the points of the curve are drawn near the diagonal.

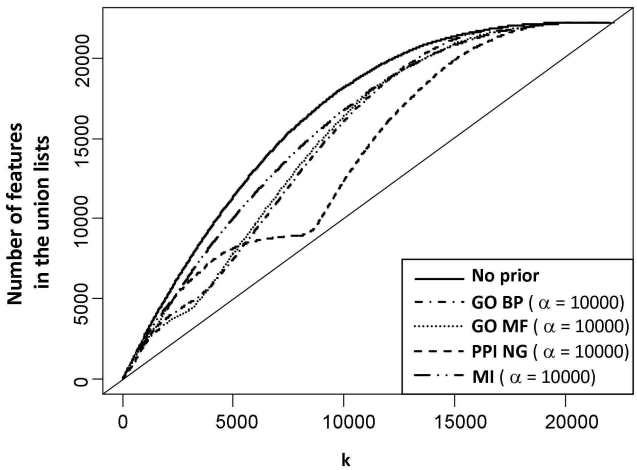

Figure 5.1: **Feature list stability.** Number of features in the union lists of length k, obtained by the standard classifier (No prior) and the integration of the best performing biological information: GO Biological Process (GO BP), GO Molecular Function (GO MF), protein-protein interactions codified by the normalized geodesic distance (PPI NG) and mutual information for gene expression data (MI).

	k	GSE2990 vs GSE3494	GSE3494 vs GSE7390	GSE7390 vs GSE7390	Mean Canberra Distance	Mean Accuracy
No prior	4182	0.95	0.94	0.94	0.94	95%
GO BP	4268	0.63	0.65	0.65	0.65	95%
GO MF	3456	0.62	0.62	0.63	0.63	94%
PPI NG	8684	0.62	0.61	0.62	0.62	96%
PPI JA	22207	0.96	0.96	0.97	0.97	95%
PPI FS	22207	0.96	0.97	0.97	0.97	96%
PPI SC	22207	0.91	0.92	0.93	0.93	95%
PE	128	0.7	0.72	0.74	0.74	96%
SP	163	0.68	0.71	0.62	0.62	95%
MI	310	0.62	0.65	0.64	0.64	93%

Table 5.3: **Canberra distance and accuracy across breast cancer datasets.** Pair-wise Canberra distance between the three breast cancer dataset at different number of features selected according to the minimum Canberra distance within datasets, using the standard classification approach without prior knowledge integration and different prior knowledge based similarity matrices. The corresponding mean value and the mean accuracy obtained across the three datasets are also reported.

Compared with the standard approach, the union lists obtained from GO BP, GO MF and PPI NG are able to improve the feature ranking, but no meaningful improvements are evident for the similarity matrix obtained using MI similarity matrix. In particular, the two GO BP and GO MF based matrices provide the most stable union lists for k around 5000 features, whereas PPI NG matrix achieves the best performance for k around 9000 features.

5.5 Discussion

In this chapter, the subject of the investigation has been the effect of using information from the biological domain into a learning process with the aim of improving its general performance with respect to the stability of predicted biomarkers. State-of-the-art machine learning methods give solutions with empirically good performance in terms of accuracy. However, if an accurate system tends to select the same biomarkers in different independent experiments, then it is more likely that the selected biomarkers are the right ones.
Gene expression data and biological prior knowledge were integrated to enhance biomarker lists stability in a classification approach. In particular, the presented analyses focused on the effect of incorporating different types of biological prior knowledge, like functional annotations, protein-protein interactions and expression correlation among genes in the learning process by evaluating biomarker list stability and classification accuracy.
Integrating prior knowledge is not an easy task since different types of information are represented in various data formats and stored in heterogeneous data structures. To do that, biological information were codified into specific pair-wise similarity measures, chosen accordingly to the type of biological information used: semantic similarities for the annotations on GO, topology-based similarity measures for PPI and correlation for gene expression data. Feature space has then been mapped into a new space in which the more similar two features are, the more closely they are mapped, since when some features are strongly correlated to each other, for example because they belong to the same biological process, then they likely have similar importance and are equally relevant for the task at hand. In other words, the weight vector obtained for a classification task should have similar values on indices relative to similar genes. Following this idea, it is possible to bias the solutions to fulfill this property. Experimental results seem to support this intuition: the proposed approach improves list stability, preserving high classification accuracy. In particular, three similarity matrices, based on

GO BP, GO MF annotations and PPI NG, are the best performers in improving list stability. The lowest gain in biomarkers list reproducibility is observed with the other matrices based on protein-protein interaction networks, although they reach convergence of the algorithm with the lower number of iterations, *i.e.* they improve class separability. In particular, the MI based matrix shows performance comparable to GO BP, GO MF and PPI NG based matrices when list stability is assessed between datasets.

The technique proposed in this chapter builds a kernel matrix from a similarity matrix, thus it can be used together with any kernel method (see [34] and references therein for a survey). In particular, it provides a standardized way to incorporate different types of biological knowledge in the kernel, with the only constraint that the information is codified by a similarity matrix.

Obtained results provide a starting point to combine similarity matrices in order to obtain even more stable biomarkers, using for example the approach proposed by Bie *et al.* [38] to combine kernels, believing that the power and potential of the proposed strategy will increase as the coverage and quality of biological databases improve.

Chapter 6

Revealing heterogeneities and inconsistencies in protein functional annotations

6.1 Background

As seen in the third chapter, each annotation in the GO has a source and a database entry attributed to it. The source can be a literature reference, a database reference or a computational evidence. All GO annotations include an evidence code to record the type of information on which the annotation is based. The most reliable annotations are those inferred directly from experimental evidence; such annotations are also important to seed the ontology so that biological functions on genes and gene products can be inferred by computational methods [154]. Even if annotations derived from direct experimental evidence are generally thought to be of higher quality than those from computational or indirect evidence, this aspect has not been yet shown robustly in the literature. As illustrated in Figure 6.1, about 99% of GO annotations are computationally derived and have not been manually curated: these are associated with the evidence code "Inferred from Electronic Annotations" (IEA). Most of these annotations come from the Gene Ontology Annotation (GOA) project at the European Bioinformatics Institute [155]. The guiding idea behind computational function annotation is the assumption that gene products with similar sequences or structures are likely to be evolutionary related and might still have similar functional roles. Electronically inferred annotations drastically extend the coverage, but at the expense of introducing a lot of noise in terms of false positive annotations and the presence

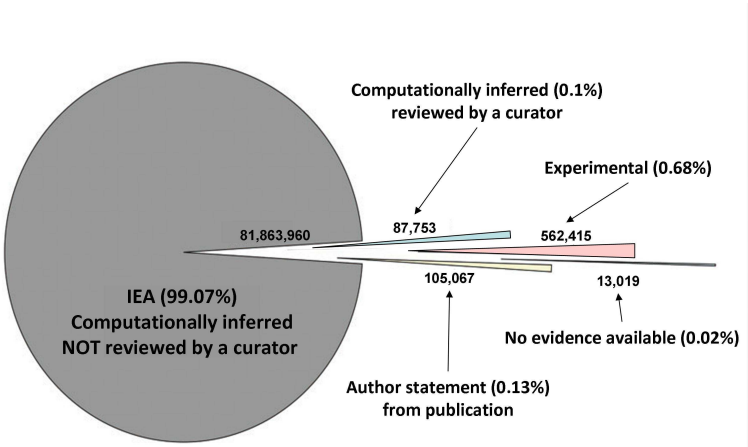

Figure 6.1: **The distribution of evidence codes among annotations in the Gene Ontology on March 2011**

of more generic annotations. This aspect characterizing GO annotation is recognized by the scientific community and the mistrust towards IEA annotations is backed by studies suggesting that annotations from available databases should be used with caution [156]. On the other hand, IEA annotations are able to provide a first insight of biological functions related to gene products and allow to extend useful analysis over some parts of the unannotated genome and proteome. However, few works considered the problem of whether or not ignoring or weighting IEA annotations in typical GO-based analysis like semantic similarity [157, 158], and a direct quantification of the extent of the problem still lack in the literature. In order to investigate this issue and to correctly interpret the accuracy and consistency of the available annotations, different approaches were recently developed. Buza *et al.* [159] provided a score of GO annotation quality based on the breadth and depth of annotations and their evidence codes, and applied it to retrieve statistics for model eukaryotes over the time. Khatri *et al.* [160] proposed a method able to discover potential inconsistencies in existing annotations and to extract implicit relationships between genes and functions in GO by the construction of a gene-function matrix. This approach, based on singular value decomposition, identifies hidden semantic links, thus providing a global assessment of the GO annotations. Considering a specific group of annotations, instead of an overall assessment, the work proposed in [161] focused on sets of functionally-related proteins and described a method that scores the degree of homogeneity of a protein set using protein-centric semantic similarity measures. However, the assessment of biological coherence of annotations often requires a detailed analysis

on which types of heterogeneities occur in the annotations, highlighting possible inconsistencies of the protein annotations in a structured and easy-to-read way. Thus, a method able both to investigate heterogeneities on GO annotations for specific user-defined pools of proteins and to address the re-annotation of protein functions is still lacking.

6.2 Data

Protein sequences were collected from UniProt Knowledgebase [162]. Only proteins with sequence length between 300 and 450 amino acids were extracted and then clustered in order to define groups of proteins characterized by high sequence similarity levels. In all, 598734 protein sequences were considered. Clustering analysis on protein sequences was performed by the algorithm CD-HIT [163], considering increasing percentage of sequence similarity: 100%, 90%, 80%, 60%, 40%. Since clustering large databases of sequences requires very time-consuming all-by-all comparisons, CD-HIT avoids many pairwise sequence alignments with a short word filter, based on the number of expected dipeptides, tripeptides and etc. which are shared between two proteins with known sequence length. Gene Ontology directed acyclic graph (DAG) structure was retrieved from the Gene Ontology Consortium [31] website (`http://www.geneontology.org`). The *gene_ontology.obo* file version 1.2 was downloaded on March 2011. Protein-GO annotations for eukaryota were derived from the Gene Ontology Annotation (GOA) database [155] (`http://www.ebi.ac.uk/GOA/`). The GOA project of the European Bioinformatics Institute provides both electronic and manual annotations to the UniProt Knowledgebase using the standardized vocabulary of the Gene Ontology (GO). The association file (*gene_association.goa_uniprot*) was downloaded from GOA on March 2011.

6.3 Global assessment of heterogeneities of the functional annotations

A first analysis of the global level of heterogeneity in GO annotations was performed, in order to quantify to what extent heterogeneities and possible inconsistencies are present in the GO database, evaluating semantic similarities on annotations of groups of proteins sharing similar functions. Secondly, a quantitative assessment of some GO properties (*e.g.* evidence code) related to hetero-

geneous annotations was performed, in order to give an overview of the current status of GO annotations which is useful to correctly interpret this kind of biological information in other bioinformatics applications, like data integration. In particular, the presented data analysis was carried out according to the following steps:

1. For each pool of proteins, grouped according to their sequence similarity, semantic similarities both between pairs of GO terms and between pairs of proteins were computed.

2. Quality Threshold clustering [164], based on semantic similarities between proteins, was applied to each group of proteins to estimate the amount of groups characterized by possible inconsistencies in the annotations, thus globally assessing the heterogeneities in the GOA database.

3. Heterogeneous annotations were investigated in terms of distribution of the related evidence codes and specificity, *i.e.* Information Content, of the related terms.

6.3.1 Semantic similarities

An information-theoretic approach was used to analyze the semantic similarity between GO terms, based on the concept of Information Content, already described in section 5.2.5. In order to semantically compare two GO terms t_1 and t_2, Lin's similarity measure [130] was used (Equation 5.7). Lin's method generates normalized similarity values in the range [0,1] and reflects how close the terms are to their common ancestor rather than just how specific that ancestor is. If proteins are well annotated near the root of the ontology, semantic similarities between the related GO terms are very high in order to avoid false positive heterogeneities due to shallow annotations.
To calculate the semantic similarity score between two proteins, pairwise GO term similarities were computed and then combined using the best-match average (BMA) approach [134]. Referring to Equation 5.10 and considering a pair of proteins p_1 and p_2, this method computes composite averages where each term of the first protein p_1 is compared only with the most similar term of the second protein p_2 and vice-versa. The BMA approach is able to robustly assess the global similarity between two proteins also when they are annotated to a different number of GO terms, since it considers both the GO terms they share and those where the proteins differ, but only the most similar ones are matched [165].

6.3.2 Quality threshold clustering

Quality Threshold (QT) clustering [164], based on semantic similarities among GOA annotations (Equation 5.10), was performed on each group of sequence-similar proteins. The QT clustering groups only proteins whose pairwise semantic similarities exceed a given user-defined quality threshold. The advantages of this method are that firstly it does not require any a priori specification on the number of clusters and secondly it does not force proteins that are dissimilar to others to be included in any cluster, thus highlighting possible outliers.
The algorithm was applied considering a grid of similarity thresholds (0.1, 0.15, 0.2,..., 0.9). For a given threshold, each group of proteins was classified according to the QT clustering output into four cases, summarized in the following:

- One Cluster (OC): proteins are all grouped into a single semantically homogeneous cluster;
- One Cluster plus Outliers (OC+O): part of the proteins are grouped into one semantic cluster, with possible protein outliers, *i.e.* proteins which are significantly dissimilar to all the others;
- More Clusters plus Outliers (MC+O): proteins are grouped into more than one semantic cluster, with possible protein outliers;
- Only Outliers (OO): proteins show pairwise semantic similarities all below the quality threshold; therefore, no clusters are identified and all proteins are considered as outliers.

Groups of proteins classified as OC+O, MC+O or OO were considered as non homogeneous with respect to GOA annotations.

6.3.3 Investigating GO properties on heterogeneous annotations

To further investigate heterogeneities in the annotation, the groups of proteins annotated as OC+O, MC+O and OO were considered separately:

- the proteins in OC+O or MC+O that are somehow clustered with other proteins, *i.e.* proteins that are not outliers;
- all the proteins in OO and the proteins in OC+O, MC+O that are outliers.

Considering these two groups, heterogeneties on GO annotations, which were identified by the clustering algorithm, were analyzed, comparing these two groups of proteins in terms of both type (*i.e.* the evidence code) and specificity (*i.e.* the Information Content) of the related GO annotations. In particular, percentages of experimental, computational and author statement annotations were calculated and the distributions of IC values of the corresponding GO terms were compared, focusing on the comparison between the first two types of evidence codes.

6.4 Functional map of heterogeneities for specific groups of proteins

An algorithm was developed to efficiently organize information on GO annotations and to highlight in an intuitive and easily interpretable way unexpected and hardly traceable heterogeneities in specific groups of sequence-similar proteins, or in groups of proteins which are clustered according to a different criterion. The algorithm performs two agglomerative hierarchical cluster analysis: the first applied to the GO terms and the second to the annotated proteins using the similarity measures based on Lin's formula and BMA approach, respectively. For each protein, only the most specific GO terms, *i.e.* those terms which are not ancestors of terms where the protein is annotated, were used in the computation of the semantic similarities. To apply the clustering algorithm, one minus the similarity value was considered as the distance measure. Average linkage method was used to update the similarity matrix. Results are organized into a functional map (see Figure 6.2) coding three different types of information represented by:

- A matrix with as many columns as the number of proteins analyzed and as many rows as the number of GO terms, where each cell (i, j) is coloured if the protein j is annotated with the GO term i, with a colour gradation representing the IC (Equation 5.4) of GO term i;
- A dendrogram for the protein-centric clustering;
- A dendrogram for the GO-centric clustering.

Although clustering analysis accounts only for the most specific GO terms of each protein, the functional map displays with coloured cells all GO annotations which are present in the database, thus preserving the original annotations. If the functional map is fully-coloured, then all the proteins share the same GO terms

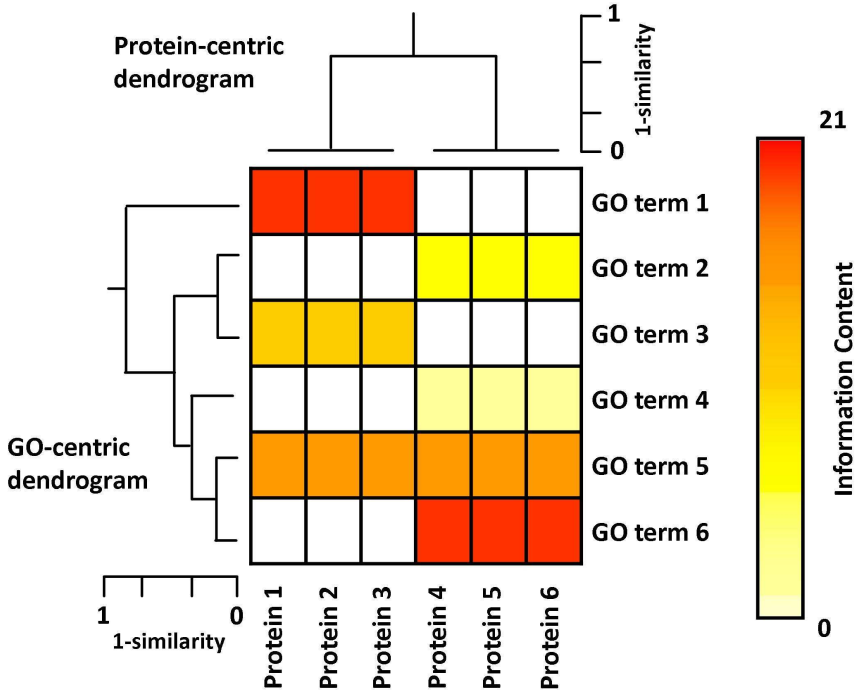

Figure 6.2: **Functional map for a group of proteins.** A colour-coded matrix is displayed, with gradations of the coloured cells ranging between light yellow to bright red for increasing IC values. Results of the two hierarchical cluster analysis are represented by the two dendrograms, displayed on the left side (GO-centric) and on the top (protein-centric) of the matrix.

and therefore they are homogeneous with respect to GO annotations. On the other hand, the presence of one or more blank cells highlights possible heterogeneites of the GO annotations.

6.5 Results

6.5.1 Global analysis on GOA database

In Figure 6.3 percentages of non homogeneous groups of proteins at different semantic thresholds are displayed for five levels of sequence similarity (100%, 90%, 80%, 60%, 40%) in both Biological Process and Molecular Function in the GOA 2011. Obviously, higher percentages of non homogeneous protein groups are observed for higher semantic thresholds and for lower sequence similarity. Considering a semantic similarity threshold equal to 0.7, the percentage of non

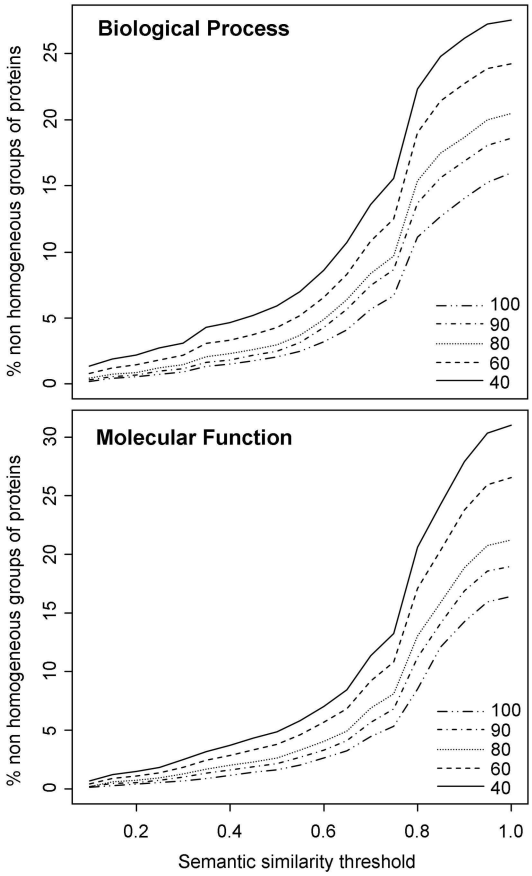

Figure 6.3: **Percentages of non homogeneous groups of proteins at different semantic thresholds for the two GO categories Biological Process and Molecular Function.** Five levels of protein sequence similarity are considered: 100% ("dashed with two dots" lines), 90% (dash dotted lines), 80% (dotted lines), 60% (dashed lines) and 40% (solid lines).

homogeneous groups of proteins ranges between 5.7% (1071 on 18928 groups) and 13.6% (3234 on 23847 groups) for Biological Process and between 4.5% (894 on 20001 groups) and 11.4% (3250 on 28608 groups) for Molecular Function. Tables 6.1 and 6.2 report how groups of proteins are partitioned into the four cases OC, OC+O, MC+O and OO (section 6.3.2) at a fixed a semantic threshold equal to 0.7 in the two GO categories. Obviously, the groups of proteins classified as OC decreases with the sequence similarity threshold, and the sum of non homogeneously annotated groups (OC+O, MC+O and OO) increases. Interestingly, the protein groups classified as OC+O or MC+O increase in number, whereas those classified as OO remain approximately constant in both Biological Process and Molecular Function.

Sequence similarity	Analyzed groups	OC	OC+O	MC+O	OO
100	18928	17857 (94.3%)	346 (1.8%)	29 (0.2%)	696 (3.7%)
90	39240	36320 (92.6%)	1506 (3.8%)	543 (1.4%)	871 (2.2%)
80	37165	34067 (91.7%)	1556 (4.2%)	735 (2.0%)	807 (2.2%)
60	30471	27186 (89.2%)	1603 (5.3%)	1018 (3.3%)	664 (2.2%)
40	23847	20613 (86.4%)	1532 (6.4%)	1176 (4.9%)	526 (2.2%)

Table 6.1: **Results of semantic clustering on Biological Process annotations.** Number of groups of proteins analyzed at different sequence similarity level and corresponding number (and percentage) of groups classified as OC, OC+O, MC+O and OO, respectively, for Biological Process and semantic similarity threshold equal to 0.7.

Sequence similarity	Analyzed groups	OC	OC+O	MC+O	OO
100	20001	19107 (95.5%)	256 (1.3%)	17 (0.1%)	621 (3.1%)
90	44220	41698 (94.3%)	1302 (2.9%)	303 (0.7%)	917 (2.1%)
80	42911	39954 (93.1%)	1514 (3.5%)	575 (1.3%)	868 (2.0%)
60	35965	32657 (90.8%)	1664 (4.6%)	920 (2.6%)	724 (2.0%)
40	28608	25358 (88.6%)	1592 (5.6%)	1087 (3.8%)	571 (2.0%)

Table 6.2: **Results of semantic clustering on Molecular Function annotations.** Number of groups of proteins analyzed at different sequence similarity level and corresponding number (and percentage) of groups classified as OC, OC+O, MC+O and OO, respectively, for Molecular Function and semantic similarity threshold equal to 0.7.

Figure 6.4 shows the percentage of annotations of proteins belonging to these two groups, annotated with different evidence code categories: experimental (EXP), computational method (COMP), author statement from publication (AUTH) and no evidence available (NOEV); the average IC for each category is also shown. The percentages shown in Figure 6.4 correspond to groups of proteins characterized by 80% of sequence similarity. Results for different level of sequence similarity are similar. As expected, the highest percentages are associated to computational annotations (COMP), which include also those electronically inferred. Interestingly, when considering the protein outliers pie charts, the percentage of experimental (EXP) and author statement annotations (AUTH) significantly increases with respect to the one observed for the proteins belonging to a cluster. This is more evident in Biological Process category, where percentages of EXP and AUTH

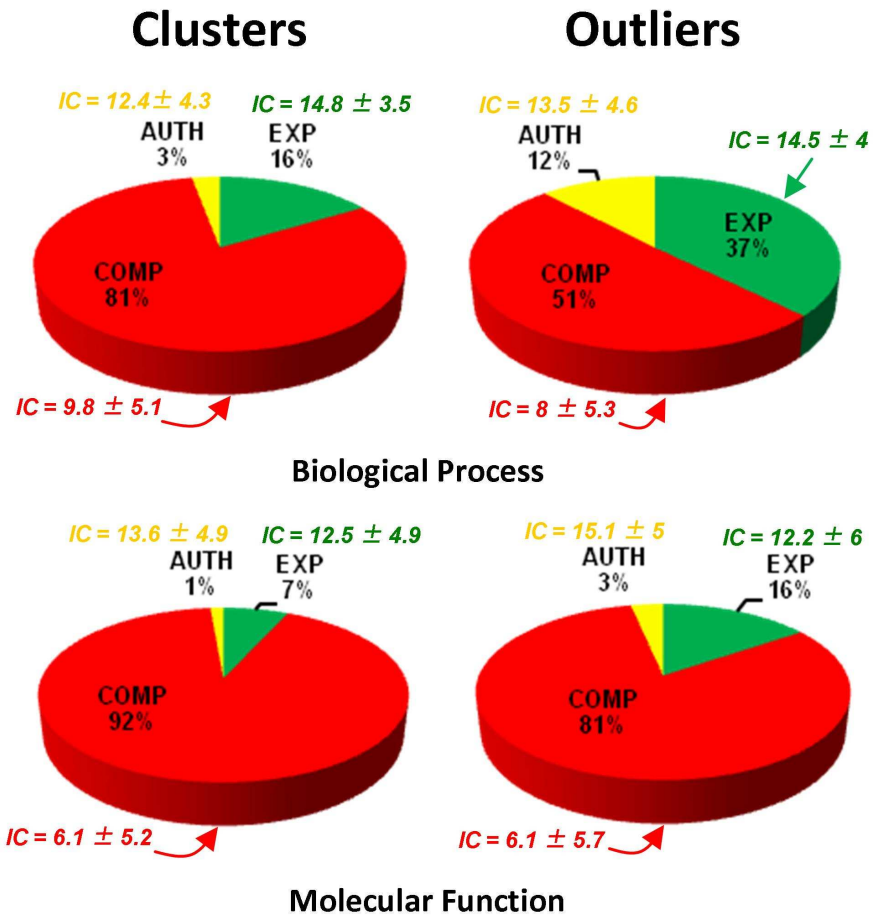

Figure 6.4: **Percentages of annotations with different evidence codes and their average IC.** The four pie-charts shows results for both BP (upper panel) and MF (lower panel) annotations of proteins belonging to groups annotated as non homogeneous. Pie-chiarts on the left show those proteins in OC+O or MC+O which are somehow clustered with other proteins; pie-charts on the right show all proteins in OO plus proteins in OC+O, MC+O that are outliers. For each category the average IC is also shown.

present a two- and four-fold increase, respectively.
Considering the average IC for each annotation category, GO terms of proteins annotated as EXP show an IC value greater than GO terms of proteins annotated as COMP. In particular, considering Molecular Function category, the average IC doubles from COMP to EXP. This seems related to the different IC content of proteins annotated with different evidence code, rather than with an intrinsic difference of IC content of protein outliers with respect to proteins belonging to

a cluster (Figure 6.5).

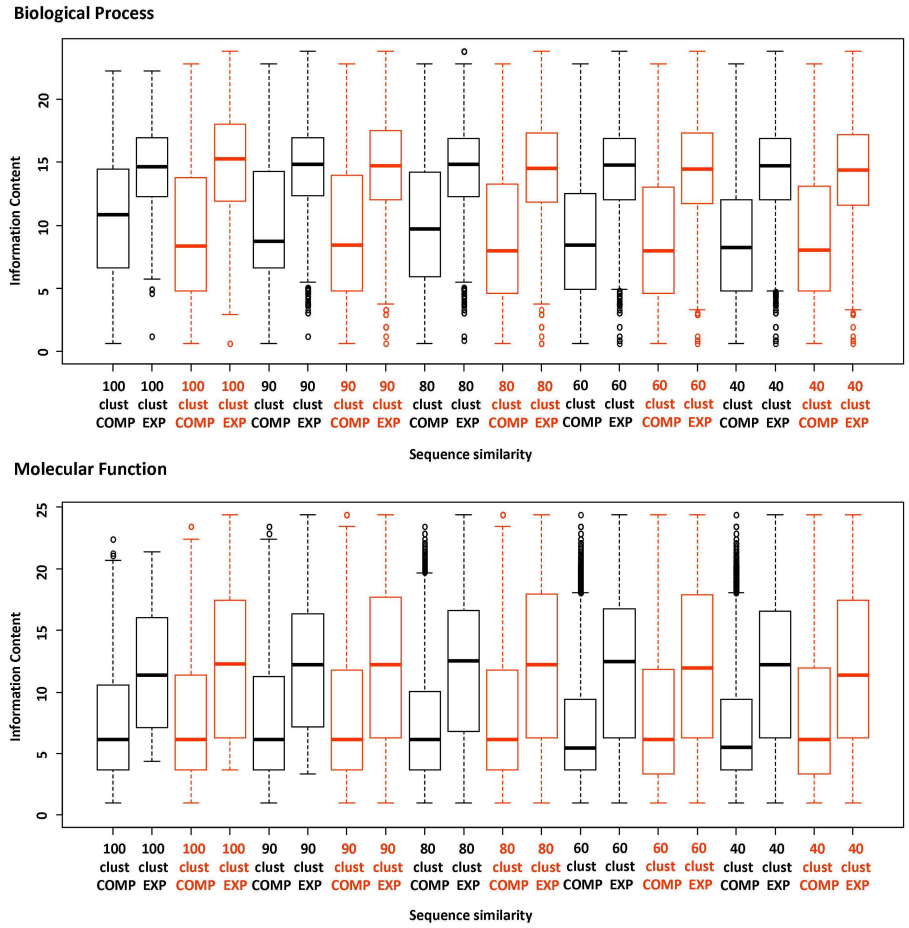

Figure 6.5: **Boxplots of the IC of GO terms for both computational (COMP) and experimental (EXP) annotations.** Results are displayed for proteins either clustered with other proteins (in black) or that are outliers (in red). Different sequence similarity levels were considered to define groups of proteins.

6.5.2 Functional map on GOA annotations: biological examples

Functional maps of many different groups of proteins were analyzed. In particular, here results of four representative groups belonging to classes OC+O, MC+O and OO are shown. These examples, depicted in Figure 6.6, have been chosen among all available in order to illustrate in detail possible interesting cases of heterogeneities in the annotation.

Examples shown in panels (a) and (b) of Figure 6.6 are representative of the

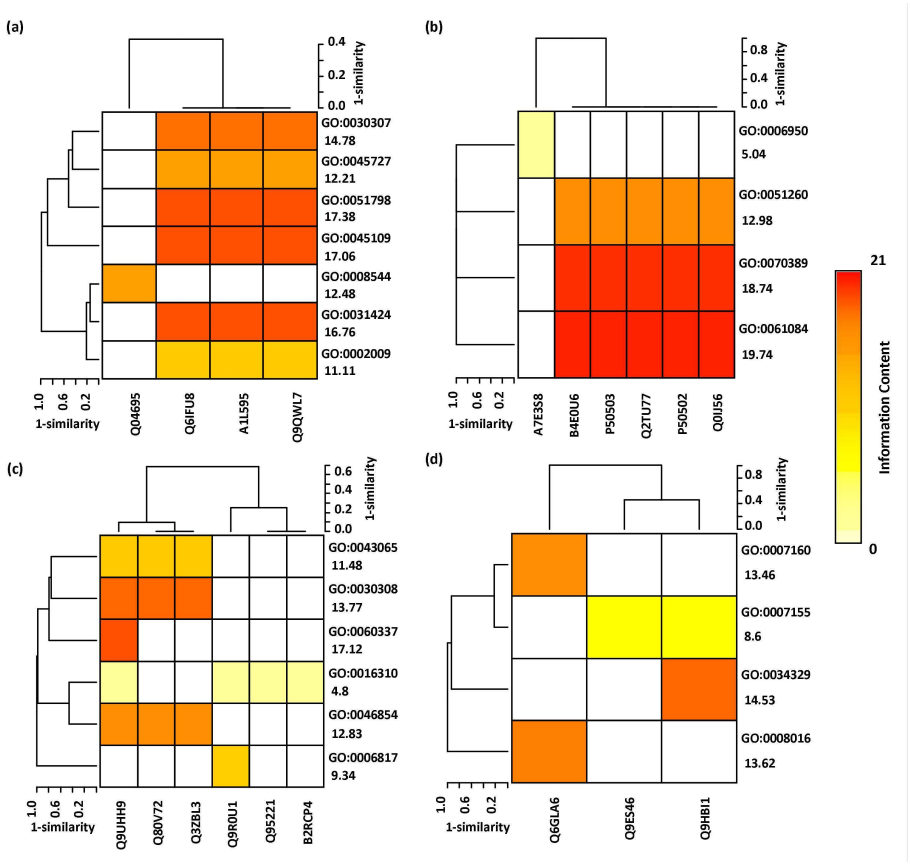

Figure 6.6: **Examples of the functional map on GOA annotations.** Panels (a) and (b) shows two examples of proteins grouped into a single cluster with the exception of one protein outlier. Two possible cases can occur: (a) the GO term associated to the protein outlier is highly similar to at least one of the other GO terms; (b) the GO term associated to the protein outlier shows low semantic similarities with respect to all the other GO terms. Panels (c) and (d) shows two examples of proteins grouped in two clusters or with respect to all the other proteins, respectively. IC values are reported below each GO term identifier displayed in the functional map.

case OC+O, *i.e.* of proteins grouped into a single cluster by the protein-centric clustering but with some protein outliers. Here, two possible cases can occur:

1. Panel (a): the outlier is annotated to a GO term very similar to at least one of the GO terms on which the clustered proteins are annotated. In the example, the GO term GO:0008544 of the protein outlier Q04695 is similar to GO terms GO:0031424 and GO:0002009. A possible interpretation can be that the protein is a missing annotation of these last two GO terms and,

in principle, could be included as their annotation. Consistently, the three proteins Q6IFU8, A1L595, Q9QWL7 could be included as annotations of the GO term GO:0008544.

2. Panel (b): the outlier is annotated to a GO term with low semantic similarity with respect to all the others. In the example, the protein A7E3S8 is annotated only to the GO term GO:0006950, suggesting that the GO term could be considered inconsistent with respect to the annotations of the analyzed group of proteins.

It is important to evidence that, in both cases, the possible interpretation of the isolated GO term is also influenced by the specificity of the molecular function or the biological process represented by the GO terms of proteins outliers: an inconsistent GO term with low IC can be neglected, whereas one with high IC needs further consideration.
If the group of proteins is classified into the cases MC+O or OO, the corresponding functional map is more heterogeneous. In these situations, the protein-centric dendrogram gives information on the number of clusters in which proteins can be divided. For instance, in the example reported in Figure 4 panel (c), the functional map allows us to identify two clusters, since proteins Q9R0U1, Q95221 and B2RCP4 are annotated to only two GO terms which are dissimilar to the GO terms where the other proteins are annotated and are characterized by low IC values. Interestingly, Q9UHH9 is the only protein annotated to the GO term GO:0060337, characterized by the highest IC value.
In the functional map reported in Figure 4 panel (d), the three proteins can be considered as three outliers, since they show low pairwise semantic similarity. For instance, the protein Q6GLA6 is annotated to two GO terms which are not shared by the other two proteins. In particular, the term GO:0008016 shows low semantic similarity to the other GO terms, as well as for the term GO:0034329, where only the protein Q9HBI1 is annotated. The only GO term with two annotated proteins is characterized by the lowest IC value.
These examples show how the proposed functional map, which combines three types of information (protein-centric similarity, GO-centric similarity, Information Content), is able to easily highlight non homogeneities in the annotations. In particular, dendrograms resulting from the two hierarchical clustering performed on both GO terms and proteins are useful to identify the main inconsistencies affecting GO annotations, whereas the Information Content is useful to assign different degrees of heterogeneities, helping the user to efficiently handle missing

or inconsistent annotations. Of note is that the biological interpretation of the heterogeneities of examples in panels (c) and (d) requires a deeper investigation than the others, since the heterogeneity is more complex.

6.6 Discussion

The identification of heterogeneities and possible inconsistent annotations affecting biological databases has become an important issue as databases increasingly rely on automated techniques for annotation. In particular, for semantically structured databases such as Gene Ontology, relationships between the terms reflect the associations between gene products and several measures have been proposed to assess semantic similarities among GO terms and the coherence between gene product annotations by comparing sets of different GO terms.
In this chapter, a global analysis performed on GOA database was shown to assess heterogeneities in the GO annotations. The best-match average method was adopted as protein-centric measures of semantic similarity, which compares all combinations of GO terms rather matching similar terms and could underestimate the similarity between the two proteins [165]. Results highlighted that possible inconsistencies on computational annotations are hardly traceable because the distribution of this type of annotation is homogeneous over the database, thus complicating the problem of managing biological information. Indeed, semantic clustering on GOA annotations highlighted that homogeneity is related to computational annotations on generic GO terms characterized by a low IC, whereas many heterogeneous annotations are experimental and associated to highly specific GO terms. However, the presence of non homogeneous protein groups demonstrated the need of a tool able to efficiently organize information on GO annotations and to highlight in an intuitive and easily interpretable way unexpected and hardly traceable heterogeneities on protein annotations.
To address this issue, a new method was developed to analyze heterogeneities of GO annotations. For groups of proteins sharing a high sequence similarity, the method performs two agglomerative hierarchical clustering based on semantic similarities both between GO terms and between protein annotations, respectively. The protein-centric analysis identifies proteins poorly annotated or annotated into different GO terms, but fails to localize these differences in GO database. Thus, it is useful to verify if GO terms of a group of proteins with a high semantic similarity are all semantically related or are associated to different biological processes or functions. The combination of GO-centric analy-

sis with the protein-centric approach allows investigating the presence of both heterogeneities among protein annotations and possible inconsistencies directly affecting GO terms. The chosen GO-centric semantic similarity measure (Lin's measure) is insensitive to the level of specificity of the common ancestor of the querying terms. This measure has been applied in order to consider the semantic similarity of two GO terms regardless of the information content of each term. In this way, the resulting functional map is able to separate the level of specificity of GO terms from their relationship in the GO graph, avoiding an overlap of information.

The algorithm gives as output a colour-coded matrix, where the presence of one or more not-colored cells could highlight possible inconsistencies among GO annotations. In this situation, dendrograms resulting from the two hierarchical clustering on both GO terms and proteins are useful to distinguish different cases of heterogeneities affecting GO annotations. Another important element for the interpretation of the functional map is the colour intensity (*i.e.* the Information Content), which is highly indicative of the degree of heterogeneity and helps to handle missing or inconsistent biological information, since the relevance of the isolated annotation depends on the specificity of the GO term. The resulting output can thus be used as a guidance for a better intepretation of the biological information associated to known proteins.

Conclusions

This thesis explored different aspects related to biomarker discovery analysis, addressing the problem from a "systems biology" point of view, thus considering and integrating different types of information of biological systems. In order to identify not only the best predictive genes or proteins characterizing a specific disease, but a stable and interpretable profile of the molecular alterations related to the disease, different computational aspects were explored, starting from the current methodologies currently used in the literature. A set of simulated and real datasets was used to test the methods separating training and test phases to avoid overfitting. Results were compared with and without the application of the bootstrap approach and with a simpler and more commonly used univariate method to evaluate the trade off between the complexity of the analysis pipeline and the improvements in terms of both classification performance and biomarker lists stability. A systematic improvement in selecting features with a high degree of precision and stability was observed for the bootstrap approach. However, the crucial factor affecting list stability seems to be that the classification task is under constrained. Looking for signatures connected on a pre-defined graph or belonging to a biological process may uncover various important aspects of the underlying biological mechanisms involved in the disease.
In particular, this thesis dealt with the computational aspects of integration of this available biological knowledge in the learning process, focusing on the reproducibility of interpretable biomarker lists, which is an issue of increasing interest from a both computational and clinical point of view. The application of external constraints with a biological meaning was analyzed, exploiting the biological information structure of Gene Ontology. Results showed that the generation of multiple classifiers from partitions of the dataset over the features and thus from less undetermined systems can improve the classification performance and automatically lead to more interpretable lists of gene sets belonging to known biological processes and molecular functions. Preliminary results on the reproducibility of these gene sets in different studies of breast cancer revealed the

identification of similar biological functions which are selected as significantly altered by the disease. A useful approach able to reduce the redundancy of information affecting GO annotations was applied; however, the multiple classifiers are built on pre-defined gene sets strictly depending on how genes are annotated to the GO terms. An alternative learning approach was developed, which is able to use biological information opportunely codified as a similarity matrix. The application of this approach on different types of biological knowledge, such as functional annotations or protein-protein interaction networks, showed that only some characteristic combinations of biological information - similarity metrics are able to achieve good performance in terms of stability of biomarkers lists preserving high levels of prediction accuracy. In particular, these results lead to some questions about handling biological information: many annotations are available, but only some are useful. The obtained results of a global analysis performed in GO annotations are indicative of the presence of heterogeneities among the GO annotations and confirm the need of considering the quality and the origin of annotations when inferring possible biological functions. If on the one hand high-throughput technologies are able to generate a huge amount of experimental data, on the other different types of prior information on genes and proteins can be a useful source to help solving these data-mining problems. This thesis provided an overview of problems and possible solutions for the integration of prior knowledge in the learning process, opening new future interesting developments. Beside the Gene Ontology, which is the most complete database of biological functions, also pathway databases are an interesting source of biological knowledge. The presented two integration approaches can be both extended to other biological representations: the first method can be applied on any source of information which can be characterized using a hierarchical structure, whereas a future possible development of the second proposed approach is to combine kernels derived from different types of knowledge in order to achieve a better characterization of stable biomarker lists. Finally, the functional map proposed in the analysis of GO annotations can be developed as a useful pre-processing method to select the most reliable information to be integrated into the learning process.

All these aspects, which have been extensively studied in different microarray data, can be hopefully also applied on next generation sequencing data, where problems related to the high number of features will be more and more relevant. Moreover, developments of new strategies to manage and organize biological information in biological databases open new challenges for more efficient methods to use these sources in the next years.

Appendices

Appendix A

The Transcriptional Response in Human Umbilical Vein Endothelial Cells Exposed to Insulin: A Dynamic Gene Expression Approach

A.1 Introduction

Type 2 diabetes is characterized by a two- to fourfold increased risk of cardiovascular disease. This is generally attributed to the adverse effects of hyperglycemia and oxidative stress on vascular biology. It has also been shown that patients with prediabetic conditions, such as impaired fasting glucose and impaired glucose tolerance, are at increased risk of cardiovascular disease as well [166]. From a pathophysiological standpoint, insulin-resistance initially induces a compensatory hyperinsulinemia, which carries on a proliferative effect among the cellular component of the vascular wall.

The *endothelium* is the thin layer of cells that lines the interior surface of blood and lymphatic vessels, forming an interface between circulating blood and lymph in the lumen and the rest of the vessel wall. The cells that form the endothelium are called *endothelial cells*. Endothelial cells in direct contact with blood are called vascular endothelial cells whereas those in direct contact with lymph are known as lymphatic endothelial cells. Vascular endothelial cells play a major role in maintaining cardiovascular homeostasis. In addition to providing a

physical barrier between the vessel wall and lumen, the endothelium secretes a number of mediators that regulate platelet aggregation, coagulation, fibrinolysis, and vascular tone. Endothelial cells secrete several mediators that can alternatively mediate either *vasoconstriction*, such as endothelin-1 and thromboxane A2, or vasodilation, such as nitric oxide (NO), prostacyclin, and endothelium-derived hyperpolarizing factor.

The first step of the adverse sequence of events that leads to the atherosclerotic process is thought to be "*endothelial dysfunction*" [167]. This term refers to a condition in which the endothelium loses its physiological properties: the tendency to promote vasodilation, fibrinolysis, and anti-aggregation. In diabetes chronic hyperinsulinemia contributes to the instability of the atherosclerotic plaque and stimulates cellular proliferation through the activation of the MAP kinases, which in turn regulate cellular proliferation. However, it is not known whether insulin itself could increase the transcription of specific genes for cellular proliferation in the endothelium. Hence, the characterization of transcriptional modifications in endothelium is an important step for a better understanding of the mechanism of insulin action and the relationship between endothelial cell dysfunction and insulin resistance.

Considering microarray experiments, in pre/post stimulus studies in which the transcriptional response is monitored at one specific time instant after a prolonged insulin exposure, genes showing a transient response followed by a return to the pre-stimulus expression or a systematic, but small in magnitude, change in the expression, are likely to be missed [168]. On the opposite, monitoring the dynamic response using more than one time samples after the stimulus allows detecting these genes as differentially expressed and provides a description of the transcriptional expression patterns of the response. Transient behavior might be characteristic, and, if common to a number of genes associated to the same functional group, might give insight into the function performed by the gene circuitry. The aim of this study is to exploit the potential of a dynamic study to investigate the dynamic transcriptional response of endothelial cells following insulin stimulation, integrating temporal profiles of genes representing the effect of insulin with the functional information of GO database on the main molecular functions characterizing the underlying biological mechanisms. This is the first systematic study in the literature monitoring transcriptional response to insulin in endothelial cells, in a time series microarray experiment.

A.2 Materials and Methods

To distinguish between insulin effect and other processes that take place in the cell simultaneously, but are not induced or inhibited by insulin, treated cells were compared with control cells. Experiments were carried out on human umbilical vein endothelial cells (HUVECs). Samples were collected at times 0, 40, 100, 200, 340, 440 min. Time 0 was cultured and harvested in duplicate so to have a complete experimental replicate of time 0 sample. Affymetrix chips were used in the experiments. Preprocessing steps such as background subtraction, probe cell normalization and expression level calculations, were performed using quantile normalization and Robust Microarray Analysis (RMA) software [84]. Data are accessible through GEO Series accession number GSE21989. High-level data analysis was carried out in a pipeline as shown in Figure A.1.

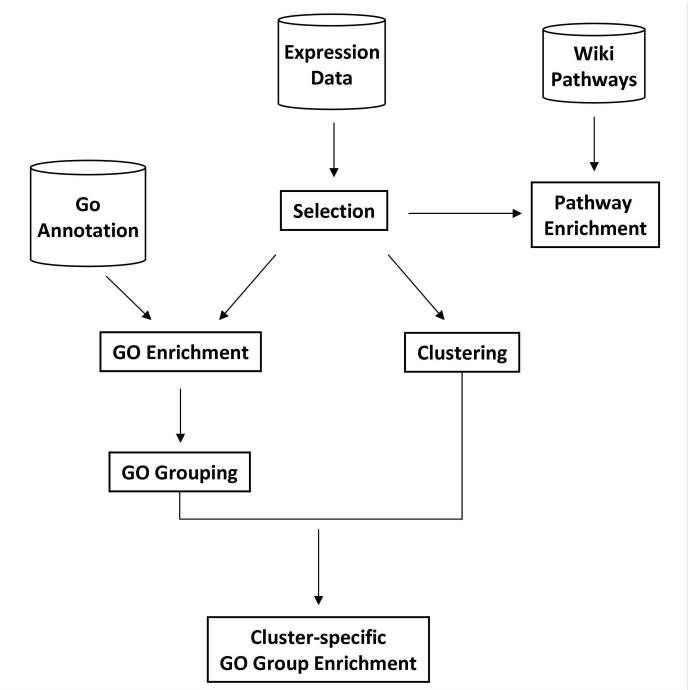

Figure A.1: **Pre-processed Affymetrix data analysis pipeline.** A selection method was applied to identify differentially expressed genes. Selected genes were clustered according to their time expression profile; significantly associated pathways and GO terms were identified through enrichment analysis. The enriched GO terms were grouped into different functional categories. For each cluster and for each GO category, GO enrichment based on Fisher's Exact Test was calculated.

Selection

Differentially expressed genes were selected using the method described in [168] that calculates the area of the region bounded by the treated-minus-control expression profile and assigns a p-value to each gene by evaluating the significance of this area against the null hypothesis. The two replicates available at time zero were used to derive the experimental error distribution at different intensity expression values and, consequently, the null hypothesis distribution of the area bounded by the treated-minus-control expression profile. As already shown in [168], the method, implemented for data poor conditions, is quite robust to random oscillation, and help diminishing both false positive and false negative rates. In order to account for multiple testing, the significance level was corrected according to a false discovery rate (FDR), *i.e.* the number of false positives divided by the number of selected genes, of 0.05.

GO and pathway enrichment

Genes were annotated according to molecular functions of Gene Ontology (GO) database, using NetAffx database. Enrichment analysis was performed based on a strategy similar to the "elim" method described in [111]. GO terms (related to different molecular functions) were grouped into levels according to the percentages of selected genes: namely, level 1 corresponds to GO terms with at least 98%-100% of their annotations selected, level 2 to the range 96%-98%, etc. Starting from level 1, the algorithm visited each level and for each GO term performed Fisher's Exact Test that assigns a p-value representing the probability that the observed number of selected genes annotated to the GO term could have resulted from random sampling. If in the visited level a GO term has a p-value below a significance level α, then the corresponding genes were removed from the annotation of GO terms having lower percentages, in order to penalize their p-value. In this way, the number of enriched GO terms was kept low, still maintaining a high significance level. Since this test was applied to a large number of GO terms, the significance level α for the calculated p-values was empirically set to 0.0025. To identify the most enriched pathways, selected genes were also annotated to WikiPathways (`http://www.wikipathways.org`) using NetAffx database. Enrichment analysis of pathways was performed using Fisher's Exact Test.

GO grouping

To obtain a more synthetic annotation, the enriched nodes directly connected by a path in the GO graph were grouped together in the same functional cluster.

Each GO group, thus characterized by an isolated sub-graph of siblings or ancestors terms, was labeled with the most general of these terms.

Clustering
To identify the main temporal expression patterns in response to insulin stimulus, treated-minus-control expression profiles of selected genes were clustered using K means clustering based on Pearson correlation. The number K of clusters was set to 7.

Cluster-specific GO group enrichment
For each cluster and for each GO group defined above, GO enrichment based on Fisher'ss Exact Test was calculated separately, so that the resulting p-value represents the probability that the observed numbers of selected genes belonging to the cluster and annotated with the GO group have resulted from random sampling. GO groups with p-value¡0.05 were considered as significantly enriched for the cluster.

A.3 Results

1715 genes were selected as differentially expressed based on their treated minus control profile, thus allowing the detection of even small but systematic changes in gene expression. Genes were clustered in 7 groups according to their time expression profile and classified into 15 functional categories that can support the biological effects of insulin, based on GO enrichment analysis (Figure A.3). The seven main temporal patterns in response to insulin stimulus were identified for treated-minus-control expression profiles as shown in Figure **??** (left panels), together with the number of genes correlated to each pattern. For each cluster the specific enrichments in the 15 different GO groups was expressed as (1-p-value), as shown in Figure **??** (right panels), so that a value close to 1 indicates an elevated significance level. These results allow to characterize the transcriptional response by remarkably different temporal profiles. For instance, cluster 1, characterized by a peak of the expression level at time 200 min, is significantly associated with GO groups B (actin binding) and C (N-terminal myristoylation domain binding), in which three genes coding for calmodulin, a protein which increases NO activity, are annotated. In terms of endothelial function, the most prominent processes affected were NADH dehydrogenase activity, N-terminal myristoylation domain binding, nitric-oxide synthase regulator activity and growth factor

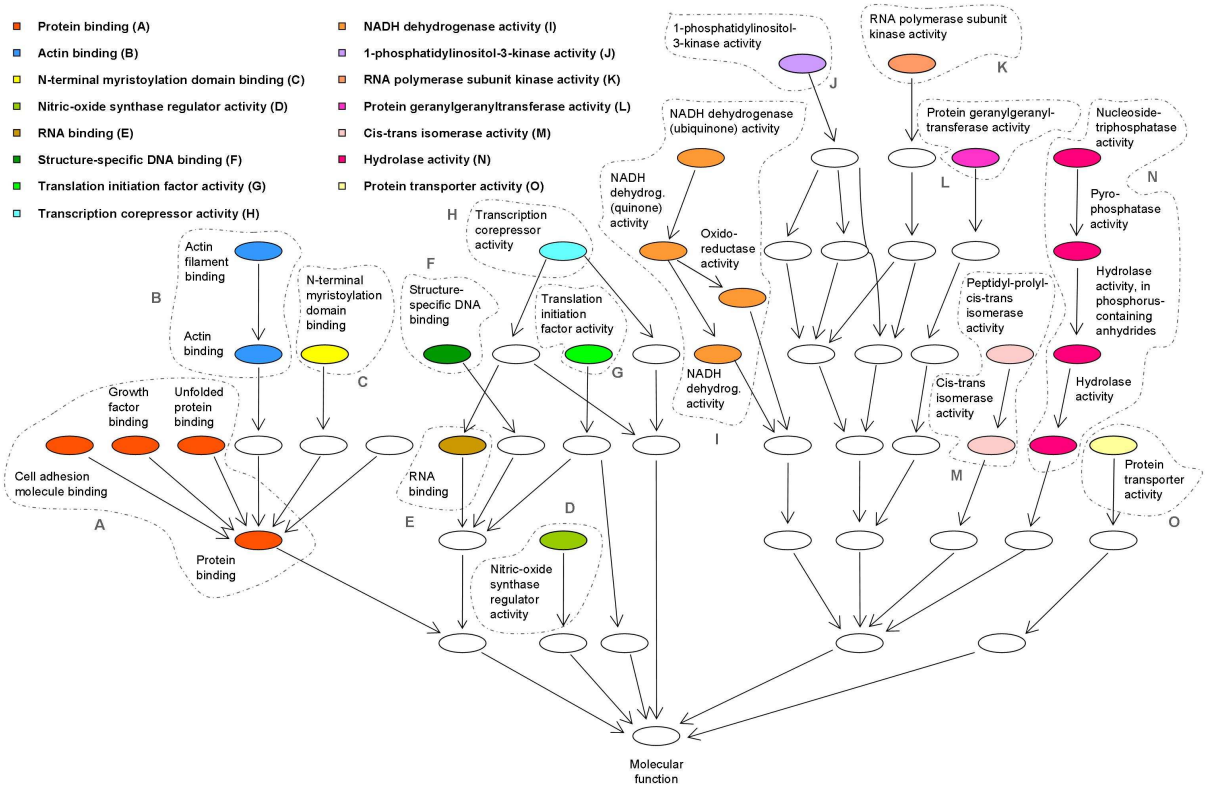

Figure A.2: **GO graph of enriched molecular function terms.** The paths of GO enriched terms are displayed; nodes directly connected by a path in the GO graph were grouped together into 15 GO main annotation groups (denoted by capital letters).

binding. Pathway-based enrichment analysis also revealed "Electron Transport Chain" as significantly enriched. Results were validated on genes belonging to this pathway, using quantitative RT-PCR.

A.4 Discussion

The objective of this study was, using DNA microarray technology, to assess the transcriptional response to insulin in endothelial cells in a time series microarray experiment and identify the main biological process and functions underlying this biological case study. To identify significant transcriptional temporal patterns in endothelial cells treated with insulin and to characterize them from a functional point of view, an ad hoc analysis pipeline was developed and applied to experimental data. In particular: 1) differentially expressed genes, selected by using a method tailored for gene expression time series in data-poor conditions,

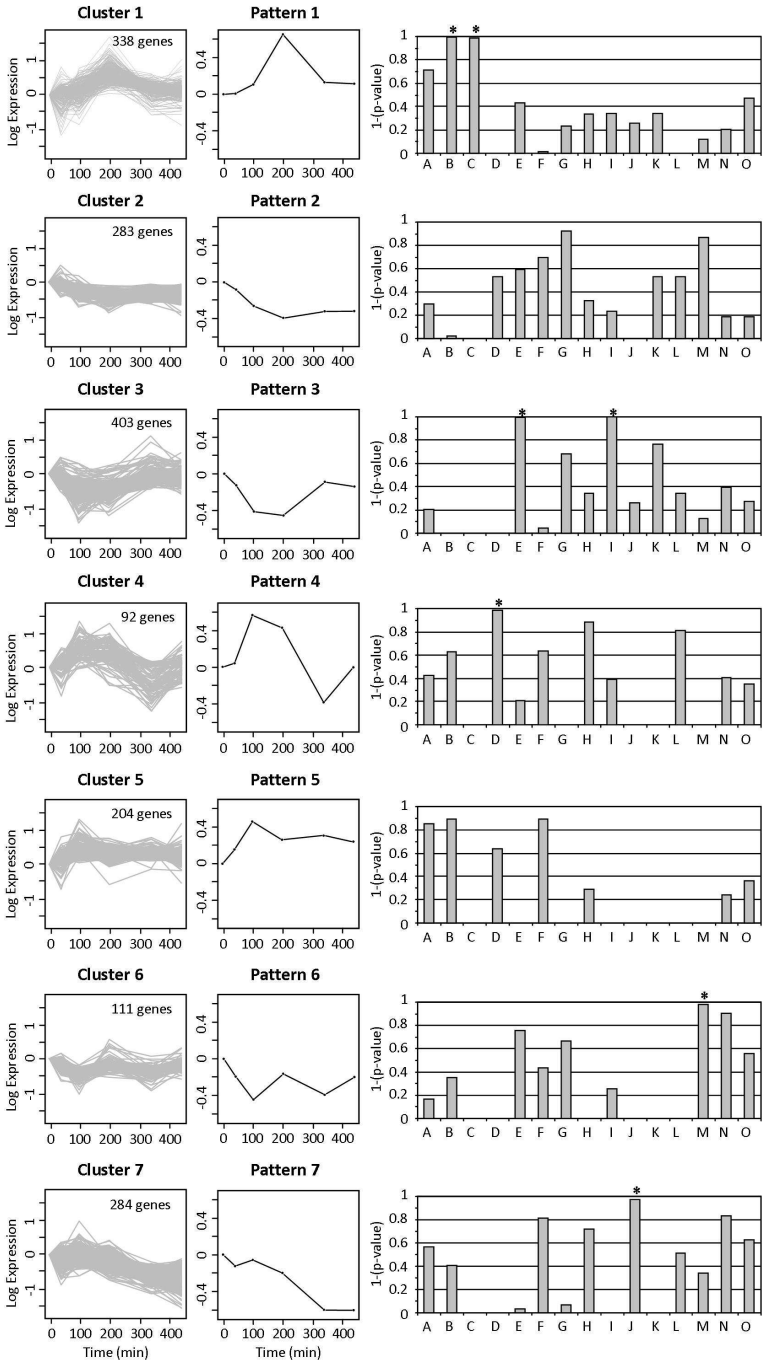

Figure A.3: **Cluster-specific GO group enrichment.** Left panels show temporal expression profiles (treated minus control) of genes belonging to each cluster, identified by K-means algorithm; middle panels represent the corresponding average temporal patterns; right panels show significance of the enrichment for each GO category (identified by capital letters) in each cluster, as 1 minus p-value. A star indicates significant GO groups (with p-value < 0.05).

were annotated according to Gene Ontology molecular functions; 2) the enriched GO terms were grouped together according to their position in the GO graph in order to obtain a more synthetic annotation and these groups were used to annotate the main temporal expression patterns identified by cluster analysis. This approach selects genes based on their dynamic gene expression profiles, thus detecting even small but systematic changes in gene expression. Then, by integrating cluster analysis and functional annotations, it gives a limited number of non-redundant functional groups.

Obtained results demonstrate that endothelial genomic response is significantly affected by elevated insulin concentrations involving a high number of genes, whose dynamic response is characterized by well-defined temporal patterns, classified into functional categories that support the biological effects of insulin. In addition, since chronic hyperinsulinemia contributes to the instability of the atherosclerotic plaque and stimulates cellular proliferation, some of the genes identified in the present work are potential novel candidates in diabetes complications related to endothelial dysfunction. More focused studies on subsets of genes and on several donors will be objective of future studies.

Appendix B

Function-based discovery of temporal patterns in insulin stimulated muscle cells

B.1 Introduction

A main objective of systems biology is the evaluation of gene/protein interactions from high-throughput time series expression data and their link to biological functions. Hence, it is important to detect the main temporal patterns characterizing the data, and associate them with functional annotation and information on differentially expressed genes. To this purpose, a first gene selection step is usually applied to the data to limit the analysis to those genes that are differentially expressed in time, and then a clustering step is applied to summarize the information using a limited number of profiles. However, both selection and clustering steps have some drawbacks. When selection of differentially expressed genes is performed it is desirable to limit the number of analyzed genes without missing the key-players, whereas selection procedures are based on a confidence threshold that controls the false positives and does not explicitly account for the false negatives. On the other hand, when using clustering to limit the number of analyzed profiles, it is desirable to obtain sufficiently homogeneous clusters, so as to summarize the signal without loosing important information, *i.e.* without missing significant changes in the pattern of expression. This aspect is related to the choice of the number of clusters, which is critical in many clustering algorithm and often let the user decide it in a qualitatively way. Moreover, the functional interpretation of these results is often performed *a posteriori*, without affecting

gene selection, and genomic databases are characterized by high redundancy of biological information, which often is not take into account.
In this study a new method is proposed that integrates selection, clustering and functional association to find the main temporal patterns associated to functional groups of significantly differentially expressed genes. The method is based on an already proposed method for gene selection [168], GO functional annotation and a new approach for gene clustering [169]. To better appreciate the enhancement of the biological interpretation gained by the proposed method, its application on real data on insulin stimulated muscle cells is shown. Skeletal muscle is responsible for about 65% of glucose disposal following a meal [170] and reduced insulin induced glucose disposal results in impaired glucose tolerance. *In vivo*, insulin plays an important role in the regulation of skeletal muscle glucose uptake and regulation of skeletal muscle protein, amino acid and fatty acid metabolism [171]. Similarly, insulin acutely stimulates protein synthesis (translation of transcripts) by activating a specific insulin-responsive protein signaling cascade [172]. Both of these responses are regulated by reversible post-translational modifications (*i.e.*, phosphorylation) of key signaling protein molecules. However, less information is available about insulin impact on gene transcription that also may affect insulin action: it is currently unknown whether insulin acutely enhances translation of genes or there is time related pattern in transcribing the genes thereby having a different level of regulation of insulin action on gene expression. Thus, the transcriptional temporal patterns remains to be fully defined.

B.2 Materials and Methods

To identify significant transcriptional temporal patterns, primary differentiated rat skeletal muscle myotubes treated with insulin were used. Samples were collected at times 0, 20, 40, 60,..., 480 minutes (every 20 minutes, for 8 hours) from both insulin treated and control cultures, for a total of 50 biological samples. Affymetrix chips were used in the experiments. Preprocessing steps such as background subtraction, probe cell normalization and expression level calculations, were performed using quantile normalization and Robust Microarray Analysis (RMA) software [84]. Data are accessible through GEO Series accession number GSE28997.
The proposed method identifies significant transcriptional temporal patterns, based on three different computational steps: 1) Gene Ranking, *i.e.* all genes are ranked according to a false discovery rate p-value reflecting the likelihood that

the gene is differentially expressed; 2) Search for the Temporal Patterns, *i.e.* each functional group is searched for temporal patterns characterizing it; 3) Selection of Differentially Expressed Genes, based on both the false discovery rate p-values and the characteristic patterns. A gene is selected as differentially expressed if it is associated to a cluster of genes: 1) sharing the same temporal transcriptional profile; 2) all annotated with the same functional term, 3) containing at least one gene with significant p-value. The output of the method is a set of clusters of differentially expressed genes, each characterized by a specific temporal pattern and by the most specific functional annotations. The three steps of the methods are described in what follows.

Gene Ranking
Genes are ranked according to false discovery rate p-values using a selection method of choice. In the case of the data analysis performed in this work, a method previously proposed was used [3] that calculates the area of the region bounded by the time series expression profile and assigns a p-value to the gene according to this area and a null hypothesis distribution, based on a model of the experimental error, to be derived from experimental replicates. The two replicates available at time zero were used to derive the experimental error distribution at different intensity expression values and, consequently, the null hypothesis distribution of the area bounded by the treated-minus-control expression profile.

Search for the Temporal Patterns
For each GO node, the algorithm searches the representative temporal patterns characterizing the transcriptional response. In particular, among the genes associated to a specific GO term, the algorithm searches for a subset of genes whose time series expression profile $X_i =< x_i(1), ..., x_i(m) >$ can be modeled by the following equation:

$$X_i = k_i \cdot P + q_i + \sum \tag{B.1}$$

where $P =< p(1), ..., p(m) >$ is the characteristic temporal expression pattern, *i.e.* a vector of m (number of time points) expression values, k_i and q_i are the gene i specific parameters and encodes the measurement error variance. The algorithm iteratively performs a gene-specific parameter identification step and a temporal pattern search step. In the first step, the parameters k_i and q_i are identified for each gene i, using weighted least squares method. A goodness-of-fit test is performed for each gene i and only genes with significant p-value are kept in the cluster. In the second step, P is estimated at each sampling time, using again

weighted least squares, but considering as data the k_i and q_i of the genes belonging to the cluster and estimated at the previous step. All the n genes being analyzed go again through the first step, so to identify new k_i and q_i and re-define the cluster membership based on the newly estimated pattern P. All the procedure is reiterated until the list of genes in the cluster does not change or a maximum number of iterations is reached. Each identified pattern is thus characterized by a cluster of genes with correlated profiles and the same annotation.
For each discovered pattern, the set of genes fitting this pattern ($fitP$) and the set of genes that do not fit it ($\neg fitP$) are defined. Only if significant, *i.e.* if it contains at least one gene with false discovery rate p-value lower than a fixed threshold, *e.g.* 0.05, $fitP$ is recorded as a cluster in the GO node under analysis. The procedure is then iteratively applied to $\neg fitP$, until $\neg fitP$ contains no genes or no significant patterns are discovered. Nodes are analyzed starting from the leaves of the GO graph, *i.e.* the nodes farthest from the root, which are the most specific GO terms; whenever a significant pattern is identified, genes correlated to the pattern are removed from all the ancestors of the node, so to avoid redundancy and annotate genes with the most specific available biological information, analogously to what has been proposed in [111]. Conversely, genes correlated to a pattern are not removed from the sibling nodes.

Selection of Differentially Expressed Genes
A gene is considered significantly differentially expressed if it has a significant false discovery rate p-value, *i.e.* lower than 0.05 (even if it is not associated to any pattern) or it is associated to a pattern P. Intuitively, since a group of genes associated to a pattern contains at least one gene with significant false discovery rate p-value, all genes in the group are likely to be differentially expressed since they are highly correlated to the same temporal pattern and share the same functional annotation.

B.3 Results

The ability of the method to identify groups of genes belonging to the same pattern was assessed on synthetic data (100 datasets of 120 profiles characterized by six different temporal patterns and 880 noisy profiles) by comparing identified to simulated clusters. The method shows high precision (96%) in detecting temporal patterns and improves selection performance by decreasing the number of false negative, maintaining constant the number of false positive (in average,

on 100 simulations the number of false negatives diminishes from 11% to 9% in correspondence of a constant false discovery rate of 5%).
Applied on insulin stimulated muscle cells, 326 genes were selected as differentially expressed and clustered into 12 different clusters, each characterized by a specific expression pattern. Figure B.1 shows in red the average differential (treated minus control) expression profiles of the genes in the different clusters; the number of genes in each cluster and their differential expression profile (in gray) is also reported. The identified clusters show patterns that includes a slow and gradual

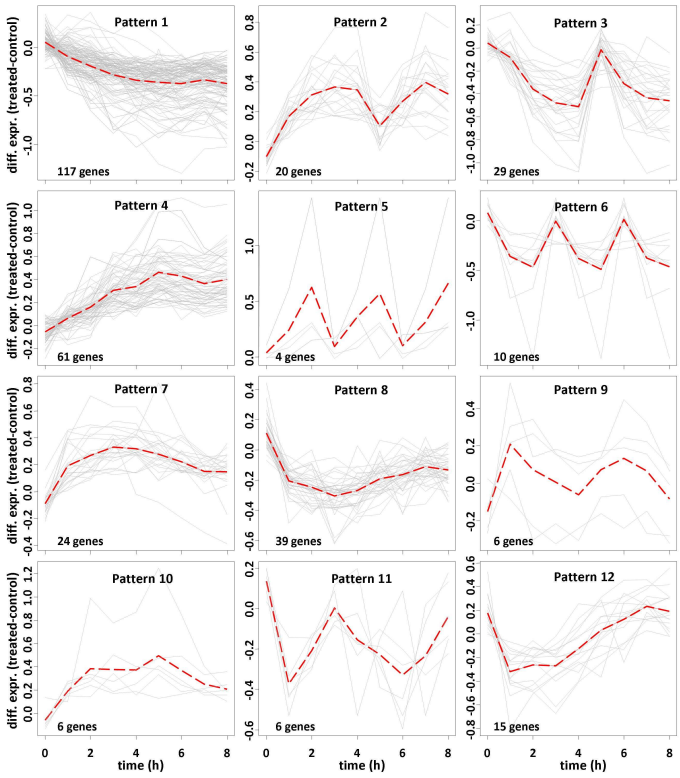

Figure B.1: **Expression profile of genes selected as differentially expressed, clustered in groups of genes sharing the same temporal patterns.** The average differential expression profiles (treated minus control) of the genes in the different clusters is shown in red; the number of genes in each cluster and their differential expression profile (in gray) is also reported for each cluster.

decrease in gene expression, a gradual increase in gene expression reaching a peak at about 5 hours and then reaching a plateau or an initial decrease and other different variable pattern of increase in gene expression over time.
Approximately 20% of the genes that were differentially expressed were identified as belonging to the insulin signaling pathway. To obtain a more synthetic annota-

tion, the GO nodes directly connected by a path in the GO graph were grouped, thus obtaining 13 GO groups plus the group of genes annotated to insulin signaling. Each GO group, thus characterized by an isolated sub-graph of siblings or ancestors terms, was labeled with the most general of these terms. Genes belonging to patterns 1 and 4, which are also the most numerous, are annotated to many different GO groups. Patterns 5 and 10, on the opposite, are characterized by genes belonging to a single GO group: nucleotide binding for pattern 5, RNA binding for pattern 10. Patterns 9, 11 and 12 are directly annotated to insulin signaling pathway, whereas patterns 2, 3, 6, 7 and 8 are in an intermediate situation, with genes annotated to a number of GO groups ranging from 3 to 6.

B.4 Discussion

The aim of the present work is to exploit the potential of a dynamic study to investigate the transcriptional response of skeletal muscle cells during acute insulin stimulation. To identify significant transcriptional temporal patterns in muscle cells treated with insulin and to characterize them from a functional point of view, a new analytical method was proposed applied to experimental data. This method aims at overcoming some drawbacks of the conventional analysis approach based on selection of differentially expressed genes, clustering and functional GO annotation. The new approach 1) improves selection of differentially expressed genes by diminishing the number of false negatives while maintaining constant the false discovery rate, *i.e.* the number of false positives divided by the number of selected genes; 2) clusters genes with the same transcriptional pattern without requiring the user to fix the number of clusters and 3) automatically annotates these clusters with the most specific GO terms, avoiding redundancy of the information.

The new method allows identifying characteristic dynamic responses to insulin stimulus, common to a number of genes and associated to the same functional group. The results demonstrate that insulin treatment elicited 12 different clusters of gene transcript profile supporting a temporal regulation of gene expression by insulin in skeletal muscle cells. Applied on GO annotations, the method is able to improve the biological interpretation and is well suited as pre-processing step in summarizing the information content with smoothed temporal profiles.

Bibliography

[1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct 20 1995.

[2] C. Sotiriou and L. Pusztai. Gene-expression signatures in breast cancer. *N Engl J Med.*, 360(8):790–800, Feb 19 2009.

[3] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 15 1999.

[4] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, Jan 31 2002.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Statist Soc B*, 57(1):289–300, 1995.

[6] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc.*, 97(457):77–87, 2002.

[7] D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander. Class prediction and discovery using gene expression data. In *Proc. Fourth Annual Intern Conf Comput Molecular Biology*, RECOMB '00, pages 263–272. ACM Press, 2000.

[8] C. S. Yu, Y. C. Chen, C. H. Lu, and J. K. Hwang. Prediction of protein subcellular localization. *Proteins*, 64(3):643–651, Aug 15 2006.

[9] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, Oct 1 2007.

[10] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther.*, 69(3):89–95, Mar 2001.

[11] F. Azuaje, Y. Devaux, and D. Wagner. Computational biology for cardiovascular biomarker discovery. *Brief Bioinform.*, 10(4):367–377, Jul 2009.

[12] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, et al. Multiple-laboratory comparison of microarray platforms. *Nat Methods*, 2(5):345–350, May 2005.

[13] L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci USA*, 103(15):5923–5928, Apr 11 2006.

[14] S. Y. Kim. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC bioinformatics*, 10:147, May 16 2009.

[15] X. Solé, N. Bonifaci, N. López-Bigas, A. Berenguer, P. Hernández, et al. Biological convergence of cancer signatures. *PLoS One*, 4(2):e4544, 2009.

[16] J. E. Larkin, B. C. Frank, H. Gavras, R. Sultana, and J. Quackenbush. Independence and reproducibility across microarray platforms. *Nat Methods*, 2(5):337–344, May 2005.

[17] J. P. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, et al. Repeatability of published microarray gene expression analyses. *Nat Genet.*, 41(2):149–155, Feb 2009.

[18] A. L. Boulesteix and M. Slawski. Stability and aggregation of ranked gene lists. *Brief Bioinform.*, 10(5):556–568, Sep 2009.

[19] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, Feb 1 2010.

[20] L. Yu, C. Ding, and S. Loscalzo. Stable feature selection via dense feature groups. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 803–811. ACM, 2008.

[21] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, Jan 15 2005.

[22] S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–492, Feb 5-11 2005.

[23] M. Zucknick, S. Richardson, and E. A. Stronach. Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Stat Appl Genet Mol Biol.*, 7(1):Article7, 2008.

[24] A. L. Boulesteix, C. Strobl, T. Augustin, and M. Daumer. Evaluating microarray-based classifiers: an overview. *Cancer Inform.*, 6:77–97, 2008.

[25] J. J. Dai, L. Lieu, and D. Rocke. Dimension reduction for classification with gene expression microarray data. *Stat Appl Genet Mol Biol.*, 5:Article6, 2006.

[26] Anne-Laure I. Boulesteix. *Dimension Reduction and Classification with High-Dimensional Microarray Data*. PhD thesis, 2005.

[27] N. Meinshausen and P. Buehlmann. Stability selection. *J R Statist Soc B*, 72(4):417–473, Jul 2010.

[28] T. Helleputte and P. Dupont. Partially supervised feature selection with regularized linear models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML'09, pages 409–416, Montreal, Jun 2009. Omnipress.

[29] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol Syst Biol.*, 3:140, 2007.

[30] J. P. Svensson, L. J. Stalpers, R. E. Esveldt van Lange, N. A. Franken, J. Haveman, et al. Analysis of gene expression using gene sets discriminates cancer patients with and without late radiation toxicity. *PLoS Med.*, 3(10):e422, Oct 2006.

[31] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, 25(1):25–29, May 2000.

[32] C. Lottaz and R. Spang. Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. *Bioinformatics*, 21(9):1971–1978, May 1 2005.

[33] F. Tai and W. Pan. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics*, 23(23):3170–3177, Dec 1 2007.

[34] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. 36(3):1171–1220, Jul 2008.

[35] V.N. Vapnik. *Statistical learning theory*. Wiley, Sep 1998.

[36] Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm. *Mach Learn.*, 37(3):277–296, Dec 1999.

[37] D. Nitsch, J. P. Goncalves, F. Ojeda, B. de Moor, and Y. Moreau. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC bioinformatics*, 11:460, Sep 14 2010.

[38] T. De Bie, L. C. Tranchevent, L. M. van Oeffelen, and Y. Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–32, Jul 1 2007.

[39] H. Frohlich, N. Speer, A. Poustka, and T. Beissbarth. GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC bioinformatics*, 8:166, May 22 2007.

[40] T. Beissbarth. Interpreting experimental results using gene ontologies. *Methods Enzymol.*, 411:340–352, 2006.

[41] A. C. Haury, P. Gestraud, and J. P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One*, 6(12):e28210, 2011.

[42] M. Y. Galperin and X. M. Fernandez-Suarez. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.*, 40(Database issue):D1–8, Jan 2012.

[43] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40(Database issue):D109–14, Jan 2012.

[44] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, and R. Herwig. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*, 39(Database issue):D712–7, Jan 2011.

[45] E. Wingender. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform.*, 9(4):326–332, Jul 2008.

[46] E. Portales-Casamar, S. Thongjuea, A. T. Kwon, D. Arenillas, X. Zhao, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 38(Database issue):D105–10, Jan 2010.

[47] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, et al. Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, 37(Database issue):D767–72, Jan 2009.

[48] C. Andorf, D. Dobbs, and V. Honavar. Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach. *BMC bioinformatics*, 8:284, Aug 3 2007.

[49] D. Devos and A. Valencia. Intrinsic errors in genome annotation. *Trends Genet.*, 17(8):429–431, Aug 2001.

[50] C. E. Jones, A. L. Brown, and U. Baumann. Estimating the annotation error rate of curated go database sequence annotations. *BMC bioinformatics*, 8:170, May 22 2007.

[51] W. R. Gilks, B. Audit, D. de Angelis, S. Tsoka, and C. A. Ouzounis. Percolation of annotation errors through hierarchically structured protein sequence databases. *Math Biosci.*, 193(2):223–234, Feb 2005.

[52] W. R. Gilks, B. Audit, D. De Angelis, S. Tsoka, and C. A. Ouzounis. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, 18(12):1641–1649, Dec 2002.

[53] C. Hadley. Righting the wrongs. *EMBO reports*, 4(9):829–831, Sep 2003.

[54] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98(9):5116–5121, Apr 24 2001.

[55] A. Buness, M. Ruschhaupt, R. Kuner, and A. Tresch. Classification across gene expression microarray studies. *BMC bioinformatics*, 10:453, Dec 30 2009.

[56] A. Dupuy and R. M. Simon. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst.*, 99(2):147–157, Jan 17 2007.

[57] C. M. Bryant, D. L. Albertus, S. Kim, G. Chen, C. Brambilla, et al. Clinically relevant characterization of lung adenocarcinoma subtypes based on cellular pathways: an international validation study. *PLoS One*, 5(7):e11712, Jul 22 2010.

[58] J. Subramanian and R. Simon. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst.*, 102(7):464–474, Apr 7 2010.

[59] P. C. Boutros, S. K. Lau, M. Pintilie, N. Liu, F. A. Shepherd, et al. Prognostic gene signatures for non-small-cell lung cancer. *Proc Natl Acad Sci USA*, 106(8):2824–2828, Feb 24 2009.

[60] V. Popovici, W. Chen, B. G. Gallas, C. Hatzis, W. Shi, et al. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.*, 12(1):R5, 2010.

[61] C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Kuffner, and R. Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19):2356–2363, Oct 1 2006.

[62] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*, 24(2):258–264, Jan 15 2008.

[63] M. Zhang, C. Yao, Z. Guo, J. Zou, L. Zhang, et al. Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics*, 24(18):2057–2063, Sep 15 2008.

[64] A. C. Haury, L. Jacob, and J. P. Vert. Improving stability and interpretability of gene expression signatures. Technical report, arXiv, 2010.

[65] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen.*, 7:179–188, 1936.

[66] D. Cai, X. He, and J. Han. SRDA: An efficient algorithm for Large-Scale discriminant analysis. *IEEE Trans Knowledge and Data Engineering*, 20(1):1–12, Jan 2008.

[67] C. J. Lin. Formulations of Support Vector Machines: A Note from an Optimization Point of View. *Neural Computation*, 13(2):307–317, 2001.

[68] Y. Sun. Iterative RELIEF for feature weighting: algorithms, theories, and applications. *IEEE Trans Pattern Anal Mach Intell.*, 29(6):1035–1051, Jun 2007.

[69] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Mach Learn.*, 46(1-3):389–422, 2002.

[70] C. Furlanello, M. Serafini, S. Merler, and G. Jurman. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC bioinformatics*, 4:54, Nov 6 2003.

[71] C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA*, 99(10):6562–6566, May 14 2002.

[72] B. Di Camillo, G. Toffolo, and C. Cobelli. A gene network simulator to assess reverse engineering algorithms. *Ann N Y Acad Sci.*, 1158:125–142, Mar 2009.

[73] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 3 2001.

[74] D. E. Featherstone and K. Broadie. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *BioEssays*, 24(3):267–274, Mar 2002.

[75] M. L. Siegal and A. Bergman. Waddington's canalization revisited: developmental stability and evolution. *Proc Natl Acad Sci USA*, 99(16):10528–10532, Aug 6 2002.

[76] B. Di Camillo, F. Sanchez-Cabo, G. Toffolo, S. K. Nair, Z. Trajanoski, and C. Cobelli. A quantization method based on threshold optimization for microarray short time series. *BMC bioinformatics*, 6 Suppl 4:S11, Dec 1 2005.

[77] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, May 2000.

[78] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello. Algebraic comparison of partial lists in bioinformatics. Technical report, arXiv, 2010.

[79] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello. Canberra distance on ranked lists. In *Proceedings of Advances in Ranking*, NIPS '09, pages 22–27, 2009.

[80] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst.*, 98(4):262–272, Feb 15 2006.

[81] L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA*, 102(38):13550–13555, Sep 20 2005.

[82] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res.*, 13(11):3207–3214, Jun 1 2007.

[83] D. S. Oh, M. A. Troester, J. Usary, Z. Hu, X. He, et al. Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. *J Clin Oncol.*, 24(11):1656–1664, Apr 10 2006.

[84] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003.

[85] S. Riccadonna, G. Jurman, S. Merler, S. Paoli, A. Quattrone, and C. Furlanello. Supervised classification of combined copy number and gene expression data. *J. Integrative Bioinformatics*, 4(3), 2007.

[86] L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol.*, 28(8):827–838, Aug 2010.

[87] P. McQuilton, S. E. St Pierre, J. Thurmond, and the FlyBase Consortium. FlyBase 101 - the basics of navigating FlyBase. *Nucleic Acids Res.*, 40(D1):D706–D714, Jan 1 2012.

[88] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, 40(Database issue):D700–5, Jan 2012.

[89] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, et al. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.*, 40(D1):D881–D886, Jan 1 2012.

[90] A. H. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, et al. Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–813, Feb 6 2004.

[91] G. D. Bader, M. P. Cary, and C. Sander. Pathguide: a pathway resource list. *Nucleic Acids Res.*, 34(Database issue):D504–6, Jan 1 2006.

[92] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39(Database issue):D691–7, Jan 2011.

[93] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow. PID: the Pathway Interaction Database. *Nucleic Acids Res.*, 37(Database issue):D674–9, Jan 2009.

[94] C. Jiang, Z. Xuan, F. Zhao, and M. Q. Zhang. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res.*, 35(Database issue):D137–40, Jan 2007.

[95] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, 40(Database issue):D841–6, Jan 2012.

[96] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, 32(Database issue):D449–51, Jan 1 2004.

[97] N. Przulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, Dec 12 2004.

[98] E. de Silva and M. P. Stumpf. Complex networks and simple models in biology. *J R Soc Interface*, 2(5):419–430, Dec 22 2005.

[99] A. L. Barabási, Z. Dezso, E. Ravasz, S. H. Yook, and Z. Oltvai. Scale-Free and Hierarchical Structures in Complex Networks. In *Modeling of Complex Systems: Seventh Granada Lectures*, volume 661, pages 1–16, Apr 2003.

[100] C. C. Friedel and R. Zimmer. Toward the complete interactome. *Nat Biotechnol.*, 24(6):614–5; author reply 615, Jun 2006.

[101] A. Ceol, A. Chatr Aryamontri, L. Licata, D. Peluso, L. Briganti, et al. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, 38(Database issue):D532–9, Jan 2010.

[102] J. De Las Rivas and C. Fontanillo. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol.*, 6(6):e1000807, Jun 24 2010.

[103] D. Nam and S. Y. Kim. Gene-set approach for expression pattern analysis. *Brief Bioinform.*, 9(3):189–197, May 2008.

[104] P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, Sep 15 2005.

[105] S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, Feb 2003.

[106] M. D. Robinson, J. Grigull, N. Mohammad, and T. R. Hughes. FunSpec: a web-based cluster interpreter for yeast. *BMC bioinformatics*, 3:35, Nov 13 2002.

[107] W. Huang da, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1):1–13, Jan 2009.

[108] L. Jacob, G. Obozinski, and J. P. Vert. Group lasso with overlap and graph lasso. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440. ACM, 2009.

[109] S. Jones, X. Zhang, D. W. Parsons, J. C. Lin, R. J. Leary, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, 321(5897):1801–1806, Sep 26 2008.

[110] S. Falcon and R. Gentleman. Using gostats to test gene lists for go term association. *Bioinformatics*, 23(2):257–258, Jan 15 2007.

[111] A. Alexa, J. Rahnenfuhrer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, Jul 1 2006.

[112] C. De Mol, S. Mosci, M. Traskine, and A. Verri. Regularization and variable selection via the elastic net. *J Comput Biol.*, 16(5):677–690, May 2009.

[113] A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970.

[114] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2003.

[115] R. Tibshirani. Regression shrinkage and selection via the lasso. *J Roy Statist Soc Ser B*, 58(1):267–288, 1996.

[116] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J Roy Statist Soc Ser B*, 67:301–320, 2005.

[117] T. Sanavia, A. Crepaldi, A. Barla, and B. Di Camillo. Gene ontology based classification improves prediction and gene signature interpretability. In *NETTAB 2011: Network Tools and Applications in Biology Workshop*, Oct 2011.

[118] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, et al. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, 39(Database issue):D1005–10, Jan 2011.

[119] H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, et al. ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, 39(Database issue):D1002–4, Jan 2011.

[120] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J. P. Vert. Classification of microarray data using gene networks. *BMC bioinformatics*, 8:35, Feb 1 2007.

[121] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, May 1 2008.

[122] M. Yousef, M. Ketany, L. Manevitz, L. C. Showe, and M. K. Showe. Classification and biomarker identification using gene network modules and support vector machines. *BMC bioinformatics*, 10:337, Oct 15 2009.

[123] H. Binder and M. Schumacher. Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC bioinformatics*, 10:18, Jan 13 2009.

[124] T. Helleputte and P. Dupont. Feature selection by transfer learning with linear regularized models. In *Lecture Notes in Artificial Intelligence*, pages 533–547, 2009.

[125] R. Herbrich, T. Graepel, and C. Campbell. Bayes point machines. *J. Mach. Learn. Res.*, 1:245–279, Sep 2001.

[126] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Trans Syst, Man, Cybern.*, 19(1):17–30, 1989.

[127] V. Pekar and S. Staab. Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In *Proceedings of the 19th Conference on Computational Linguistics, COLING-2002, August 24 - September 1, 2002 , Taipei, Taiwan*, 2002.

[128] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, Jul 1 2003.

[129] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[130] D. Lin. An information-theoretic definition of similarity. In *Proceedings 15th International Conf. on Machine Learning, 24-27 July*, pages 296–304, Madison, Wisconsin, USA, 1998. Morgan Kaufmann, San Francisco, CA.

[131] J. L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, et al. Correlation between gene expression and go semantic similarity. *IEEE/ACM Trans Comput Biol Bioinform.*, 2(4):330–338, Oct-Dec 2005.

[132] Y. Tao, L. Sam, J. Li, C. Friedman, and Y. A. Lussier. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*, 23(13):i529–38, Jul 1 2007.

[133] L. du Plessis, N. Skunca, and C. Dessimoz. The what, where, how and why of gene ontology–a primer for bioinformaticians. *Brief Bioinform.*, 12(6):723–735, Nov 2011.

[134] F. Couto, M. Silva, and P. Coutinho. Measuring semantic similarity between gene ontology terms. *Data & Knowledge Engineering*, 61(1):137–152, Apr 2007.

[135] A. L. Barabasi, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nat Rev Genet.*, 12(1):56–68, Jan 2011.

[136] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Mol Syst Biol.*, 3:88, 2007.

[137] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[138] H. N. Chua, W. K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, Jul 1 2006.

[139] Y. R. Cho and A. Zhang. Identification of functional hubs and modules by converting interactome networks into hierarchical ordering of proteins. *BMC bioinformatics*, 11 Suppl 3:S3, Apr 29 2010.

[140] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18 Suppl 2:S231–40, 2002.

[141] I. Priness, O. Maimon, and I. Ben-Gal. Evaluation of gene-expression clustering via mutual information distance measure. *BMC bioinformatics*, 8:111, Mar 30 2007.

[142] N. Gupta and S. Aggarwal. MIB: Using mutual information for biclustering gene expression data. *Pattern Recognition*, 42:2692–2697, Mar 2010.

[143] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–679, Feb 19-25 2005.

[144] S. Loi, B. Haibe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, et al. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol.*, 25(10):1239–1246, Apr 1 2007.

[145] M. Schmidt, D. Bohm, C. von Torne, E. Steiner, A. Puhl, et al. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.*, 68(13):5405–5413, Jul 1 2008.

[146] Y. Pawitan, J. Bjohle, L. Amler, A. L. Borg, S. Egyhazi, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.*, 7(6):R953–64, 2005.

[147] X. Lu, X. Lu, Z. C. Wang, J. D. Iglehart, X. Zhang, and A. L. Richardson. Predicting features of breast cancer with gene expression patterns. *Breast cancer Res Treat.*, 108(2):191–201, Mar 2008.

[148] B. J. Boersma, M. Reimers, M. Yi, J. A. Ludwig, B. T. Luke, et al. A stromal gene signature associated with inflammatory breast cancer. *Int J Cancer.*, 122(6):1324–1332, Mar 15 2008.

[149] A. V. Ivshina, J. George, O. Senko, B. Mow, T. C. Putti, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.*, 66(21):10292–10301, Nov 1 2006.

[150] C. O. Daub, R. Steuer, J. Selbig, and S. Kloska. Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data. *BMC bioinformatics*, 5:118, Aug 31 2004.

[151] A. M. Law and D. W. Kelton. *Simulation Modelling and Analysis*. McGraw-Hill, Apr 2000.

[152] H. A. Sturges. The choice of a class interval. *J Am Stat Assoc.*, 21(153):65–66, Mar 1926.

[153] A. Bisognin, A. Coppe, F. Ferrari, D. Risso, C. Romualdi, S. Bicciato, and S. Bortoluzzi. A-madman: annotation-based microarray data meta-analysis tool. *BMC bioinformatics*, 10:201, Jun 29 2009.

[154] Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol.*, 5(7):e1000431, Jul 2009.

[155] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler. The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, 37(Database issue):D396–403, Jan 2009.

[156] A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol.*, 5(12):e1000605, Dec 2009.

[157] X. Guo, R. Liu, C. D. Shriver, H. Hu, and M. N. Liebman. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 22(8):967–973, Apr 15 2006.

[158] S. Jain and G. D. Bader. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC bioinformatics*, 11:562, Nov 15 2010.

[159] T. J. Buza, F. M. McCarthy, N. Wang, S. M. Bridges, and S. C. Burgess. Gene Ontology annotation quality analysis in model eukaryotes. *Nucleic Acids Res.*, 36(2):e12, Feb 2008.

[160] P. Khatri, B. Done, A. Rao, A. Done, and S. Draghici. A semantic analysis of the annotations of the human genome. *Bioinformatics*, 21(16):3416–3421, Aug 15 2005.

[161] M. Chagoyen, J. M. Carazo, and A. Pascual-Montano. Assessment of protein set coherence using functional annotations. *BMC bioinformatics*, 9:444, Oct 20 2008.

[162] M. Magrane and UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database*, 2011:bar009, Mar 29 2011.

[163] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, Jul 1 2006.

[164] L. J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.*, 9(11):1106–1115, Nov 1999.

[165] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto. Semantic similarity in biomedical ontologies. *PLoS Comput Biol.*, 5(7):e1000443, Jul 2009.

[166] D. Kirpichnikov and J. R. Sowers. Diabetes mellitus and diabetes-associated vascular disease. *Trends Endocrinol Metab.*, 12(5):225–230, Jul 2001.

[167] A. Avogaro, M. Albiero, L. Menegazzo, S. de Kreutzenberg, and G. P. Fadini. Endothelial dysfunction in diabetes: the role of reparatory mechanisms. *Diabetes care*, 34 Suppl 2:S285–90, May 2011.

[168] B. Di Camillo, G. Toffolo, S. K. Nair, L. J. Greenlund, and C. Cobelli. Significance analysis of microarray transcript levels in time series experiments. *BMC bioinformatics*, 8 Suppl 1:S10, Mar 8 2007.

[169] B. Di Camillo, G. Toffolo, and C. Cobelli. Function-based discovery of temporal patterns in high-throughput genomic studies. In *Foundation in Systems Biology in Engineering*, Stuttgart, Germany, 2007.

[170] G. I. Shulman, D. L. Rothman, T. Jue, P. Stein, R. A. DeFronzo, and R. G. Shulman. Quantitation of muscle glycogen synthesis in normal subjects and subjects with non-insulin-dependent diabetes by 13c nuclear magnetic resonance spectroscopy. *N Engl J Med.*, 322(4):223–228, Jan 25 1990.

[171] L. S. Chow, R. C. Albright, M. L. Bigelow, G. Toffolo, C. Cobelli, and K. S. Nair. Mechanism of insulin's anabolic effect on muscle: measurements of muscle protein synthesis and breakdown using aminoacyl-tRNA and other surrogate measures. *Am J Physiol Endocrinol Metab.*, 291(4):E729–36, Oct 2006.

[172] P. H. Scott, G. J. Brunn, A. D. Kohn, R. A. Roth, and J. C. Lawrence Jr. Evidence of insulin-stimulated phosphorylation and activation of the mammalian target of rapamycin mediated by a protein kinase B signaling pathway. *Proc Natl Acad Sci USA*, 95(13):7772–7777, Jun 23 1998.