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

1

Integrative data analysis and visualisation: Introduction to critical problems, goals and challenges

Francisco Azuaje, Joaquin Dopazo

Abstract

This chapter introduces fundamental concepts and problems approached in this book. A rationale for the application of integrative data analysis and visualisation approaches is presented. Critical design, implementation and evaluation factors are discussed. It identifies barriers and opportunities for the development of more robust and meaningful methods. It concludes with an overview of the content of the book.

Key words: Biological data analysis, data visualisation, integrative data analysis, functional genomics, systems biology, design principles.

1. Data analysis and visualisation: An integrative approach

With the popularisation of high throughput technologies, and the consequent enormous accumulation of biological data, the development of a systems biology era will depend on the generation of predictive models and their capacity to identify and combine multiple information resources. Such data, knowledge and models are associated with different levels of biological organisation. Thus, it is fundamental to improve the understanding of how to integrate biological information which is complex, heterogeneous and geographically distributed.

The *analysis* (including discovery) and *visualisation* of relevant biological data patterns have been traditionally approached as independent computational problems.

Until now biological data analysis has placed emphasis on the automation aspects of tools, and relatively little attention has been given to the integration and visualisation of information and models, probably due to the relative simplicity of pre-genomic data. However, in the post-genomic era it is very convenient that, these tasks complement each other in order to achieve higher integration and understanding levels.

This book provides scientists and students with the basis for the development and application of integrative computational methods to exchange and analyse biological data on a systemic scale. It emphasises the processing of multiple data and knowledge resources, and the combination of different models and systems. One important goal is to address existing limitations, new requirements and solutions by providing comprehensive descriptions of techniques and applications. It covers different data analysis and visualisation problems and techniques for studying the roles of genes and proteins at a systems level. Thus, we have adopted a fairly broad definition for the areas of *genomics* and *proteomics*, which also comprises a wider spectrum of *"omic"* approaches required for the understanding of the *functions* of genes and their products.

Emphasis is placed on *integrative* biological and computational approaches. Such an integrative framework refers to the study of biological systems based on the combination of data, knowledge and predictive models originating from different sources. It brings together informational views and knowledge relevant to or originating from diverse organisational, functional modules.

Data analysis comprises systems and tools for identifying, organising and interpreting relevant biological patterns in databases as well as for asking functional questions in a whole-genome context. Typical functional data analysis tasks include classification, gene selection or their use in predictors for microarray data, the prediction of protein interactions, etc.

Data visualisation covers the design of techniques and tools for formulating, browsing and displaying prediction outcomes and complex database queries. It also covers the automated description and validation of data analysis outcomes.

Biological data analysis and visualisation have traditionally been approached as independent problems. Relatively little attention has been given to the integration and visualisation of information and models. However, the integration of these areas facilitates a deeper understanding of problems at a systemic level.

Traditional data analysis and visualisation lack key capabilities required for the development of a systems biology paradigm. For instance, biological information visualisation has typically consisted of the representation and display of information associated with lists of genes or proteins. Graphical tools have been implemented to visualise more complex information, such as metabolic pathways and genetic networks. Recently, more complex tools, such as *Ensembl* (Birney *et al.*, 2003), integrate different types of information, e.g. genomic, functional, polymorphisms, etc., on a genome-wide context. Other tools, such as *GEPAS* (Herrero *et al.*, 2004), integrate gene expression data as well as genomic and functional information for predictive analysis. Nevertheless, even state-of-the-art tools still lack the elements necessary to achieve a meaningful, robust integration and interpretation of multiple data and knowledge sources.

This book aims to present recent and significant advances in data analysis and visualisation that can support system biology approaches. It will discuss key design, application and evaluation principles. It will address the combination of different types of biological data and knowledge resources, as well as prediction models and analysis tools. From a computational point of view it will demonstrate: a) how data analysis techniques can facilitate more comprehensive, user-friendly data visualisation tasks; and b) how data visualisation methods may make data analysis a more meaningful and biologically-relevant process. This book will describe how this synergy may support integrative approaches to functional genomics.

2. Critical design and implementation factors

This section briefly discusses important data analysis problems that are directly or partially addressed by some of the subsequent chapters.

Over the past 8 years a substantial collection of data analysis and prediction methods for functional genomics has been reported. Among the many papers published in journals and conference proceedings, perhaps only a minority perform rigorous comparative assessment against well-established and previously tested methodologies. Moreover, it is essential to provide more scientifically sound problem formulations and justifications. This is especially critical when adopting methodologies involving, for example, assumptions about the statistical independence between predictive attributes or the interpretation of statistical significance.

Such technical shortcomings and the need to promote health and wealth through innovation represent strong reasons for the development of shared, best practices for data analysis applications in functional genomics. This book includes contributions addressing one or more of these critical factors for different computational and experimental problems. They describe approaches, assess solutions and critically discuss their advantages and limitations.

Supervised and unsupervised classification applications are typical, fundamental tasks in functional genomics. One of the most challenging questions is not whether there are techniques available for different problems, but rather "which" specific technique(s) should be applied and "when" to apply them. Therefore, data analysis models must be evaluated to detect and control unreliable data analysis conditions, inconsistencies and irrelevance. A well-known scheme for supervised classification is to generate indicators of accuracy and precision. However, it is essential to estimate the significance of the differences between prediction outcomes originating from different models. It is not uncommon to find studies published in recognised journals and conferences that claim prediction quality differences, which do not provide evidence of statistical significance given the data available and the models under comparison. Chapters 5 and 12 are particularly relevant to understand these problems.

The lack of adequate evaluation methods also negatively affects clustering-based studies (see Chapters 7, 10 and 11). Such studies must provide quality indicators to measure the significance of the obtained clusters, for example in terms of their compactness and separation. Another important factor is to report statistical evidence to support the choice of a particular number of clusters. Furthermore, in annotation-based analyses it essential to apply tools to determine the functional classes (such as Gene Ontology terms) that are significantly enriched in a given cluster (see Chapter 7)

Predictive *generalisation* is the ability to correctly make predictions (such as classification) on data unseen during the model implementation process (sometimes referred to as training or learning). Effective and meaningful predictive data analysis studies should aim to build models able to generalise. It is usually accepted that a model will be able to achieve this property if its architecture and learning parameters have been properly selected. It is also critical to ensure that enough training data are available to build the prediction model. However, such a condition is difficult to satisfy due to resource limitations. This is a key feature exhibited, for instance, by a significant number of gene expression analyses. With a small set of training data, a prediction model may not be able to accurately represent the data under analysis. Similarly, a small test dataset may contribute to an unreliable prediction quality assessment. The problems of building prediction models based on small datasets and the estimation of their predictive quality deserve a more careful consideration in functional genomics. Model over-fitting is a significant problem for designing effective and reliable prediction models. One simple way to determine that a prediction model, *M*, is over-fitting a training dataset consists of identifying a model *M'*, which exhibits both higher training prediction and lower test prediction errors in relation to *M*. This problem is of course directly linked to the prediction generalisation problem discussed above. Thus, an over-fitted model is not capable to make accurate predictions on unseen data. Several predictive quality assessment and data sampling techniques are commonly applied to address this problem. For example, the prediction performance obtained on a *validation dataset* may be used to estimate when a neural network training process should be stopped to improve generalisation. Over-fitting basically indicates that a prediction learning process was not correctly conducted due to factors such as an inadequate selection of training data and/or learning parameters. The former factor is commonly a consequence of the availability of small datasets. It is crucial to identify factors, experimental conditions and constraints that contribute to over-fitting in several prediction applications for functional genomics. This type of studies may provide guidelines to make well-informed decisions on the selection of prediction models. Solutions may be identified not only by looking into these constraints, but also by clearly distinguishing between prediction goals. A key goal is to apply models, architectures and learning parameters that provide both an accurate and robust representation of the data under consideration. Further research is needed to understand how to adapt and combine prediction methods to avoid over-fitting problems in the presence of small or skewed data problems.

Feature selection is another important problem relevant to predictive data analysis and visualisation. The problem of selecting the most relevant features for a classification problem has been typically addressed by implementing *filter* and *wrapper* approaches. Filter-based methods consist of statistical tests to detect features that are significantly differentiated among classes. Wrapper approaches select relevant features as part of the optimization of a classification problem, i.e. they are embedded into the classification learning process. Wrapper methods commonly outperform filter methods in terms of prediction accuracy. However, key limitations have been widely studied. One such limitation is the *instability problem*. In this problem variable, inconsistent feature subsets may be selected even for small variations in the training datasets and classification architecture. Moreover, wrapper methods are more computationally expensive. Instability may not represent a critical problem if the main objective of the feature selection task is to optimise prediction performance, such as classification accuracy. Nevertheless, deeper investigations are required if the goal is to assess biological relevance of features, such as the discovery of potential biomarkers. Further research is necessary to design methods capable of identifying robust and meaningful feature relevance. These problems are relevant to the techniques and applications presented in Chapters 5, 6, 12 and 13.

The area of functional genomics present novel and complex challenges, which may require a redefinition of conceptions and principles traditionally applied to areas such as engineering or clinical decision support systems. For example, one important notion is that significant, meaningful feature selection can be achieved through both the reduction and maximisation of feature *redundancy* and *diversity* respectively.

Therefore, crucial questions that deserve deeper discussions are: Can feature similarity (or correlation) be associated with redundancy or irrelevance? Does feature diversity guarantee the generation of biologically meaningful results? Is feature diversity synonym of relevance? Sound answers will of course depend on how concepts such as feature relevance, diversity, similarity and redundancy are defined in both computational and biological contexts.

Data mining and knowledge discovery consist of several, iterative and interactive analysis tasks, which may require the application of heterogeneous and distributed tools. Moreover, a particular analysis and visualisation outcome may represent only a component in a series of processing steps based on different software and hardware platforms. Therefore, the development of system- and application-independent schemes for representing analysis results is important to support more efficient, reliable and transparent information analysis and exchange. It may allow a more structured and consistent representation of results originating from large-scale studies, involving for example several visualisation techniques, data clustering and statistical significance tests. Such representation schemes may also include metadata or other analysis content descriptors. They may facilitate not only the reproducibility of results, but also the implementation of subsequent analyses and inter-operation of visualisation systems (Chapter 9). Another important goal is to allow their integration with other data and information resources. Advances mainly oriented to the data generation problem, such as the *MicroArray Gene Expression Markup Language* (MAGE-ML), may offer useful guidance to develop methods for the representation and exchange of predictive data analysis and visualisation results.

3. Overview of contributions

The remaining of the book comprises 13 chapters. The next two chapters overview key concepts and resources for data analysis and visualisation. The second part of the book focuses on systems and applications based on the combination of multiple types of data. The third part highlights the combination of different data analysis and visualisation predictive models.

Chapter 2 provides a survey of current techniques in data integration as well as an overview of some of the most important individual databases. Problems derived from the enormous complexity of biological data and from the heterogeneity of data sources in the context of data integration and data visualization are discussed.

Chapter 3 overviews fundamental concepts, requirements and approaches to a) Integrative data analysis and visualisation approaches with an emphasis on the processing of multiple data types or resources; and b) integrative data analysis and visualisation approaches with an emphasis on the combination of multiple predictive models and analysis techniques. It also illustrates problems in which both methodologies can be successfully applied, and discusses design and application factors.

Chapter 4 introduces different methodologies of text mining and the current status, possibilities and limitations offered by those methods as well as their relation with the corresponding areas of molecular biology, with particular focus on the analysis of protein interaction networks.

Chapter 5 introduces a probabilistic model that integrates multiple information sources for the prediction of protein interactions. It presents an overview of genomic sources and machine learning methods, and explains important network analysis and visualisation techniques.

Chapter 6 focuses on the representation and use of genome-scale phenotypic data which, in combination with other molecular and bioinformatic data open new possibilities for understanding and modelling the emergent complex properties of the cell. QTL analysis, reverse genetics and phenotype prediction in the new pors-genomics scenario are discussed.

Chapter 7 shows an overview on the use of bioontologies in the context of functional genomics with special stress on the most extensively used one: gene ontology. Important statistical issues related to high-throughput methodologies, such as the high occurrence of false or spurious associations between groups of genes and functional terms when the proper analysis is not performed, are also discussed.

Chapter 8 discusses data resources and techniques for generating and visualising interactome networks with an emphasis on the interactome of *C. elegans*. It overviews technical aspects of the large scale high-throughput yeast two hybrid approach, topological and functional properties of the interactome network of *C. elegans*, and their relationships with other sources such as expression data.

Chapter 9 reviews some of the limitations that tools for data management and visualization present. It introduces the UTOPIA, a project in which re-usable software components are being built and integrated closely with the familiar desktop environment to make easy-to-use visualisation tools for the field of bioinformatics.

Chapter 10 reviews fundamental approaches and applications to data clustering. It focuses on requirements and recent advances for gene expression analysis. This contribution discusses crucial design and application problems for interpreting, integrating and evaluating results.

Chapter 11 introduces an integrative, unsupervised analysis framework for microarray data. It stresses the importance of implementing integrated analysis of heterogeneous biological data for supporting gene function prediction. It explains how multiple clustering models may be combined to improve predictive quality. It focuses on the design, application and evaluation of a knowledge-based tool that integrates probabilistic, predictive evidence originating from different sources.

Chapter 12 reviews well accepted supervised methods to address questions about differential expression of genes and class prediction from gene expression data. Problems that limit the potential of supervised methods are analised with special stress on issues such as inadequate validation of error rates the non rigorous use of data sets and the failure to recognise observational studies and include needed covariates.

Chapter 13 presents an overview on inferring genetic networks by probabilistic graphical models. Different types of probabilistic graphical models are introduced and methods for learning these models from data are presented. The use of probabilistic graphical models for modelling molecular networks at different levels is also discussed.

Chapter 14 introduces key approaches to the analysis, prediction and comparison of protein structures. For example, it stresses the application of a method that detects local patterns in large sets of structures. This chapter illustrates how advanced approaches may not only complement traditional methods, but also how they may provide alternative, meaningful views of the prediction problems.

References

Birney, E., Ensembl Team (2003) Ensembl: a genome infrastructure. *Cold Spring Harb Symp Quant Biol*, 68, 213-5.

Herrero, J., Vaquerizas, J.M., Al-Shahrour, F., Conde, L., Mateos, A., Santoyo, J., Diaz-Uriarte, R., Dopazo, J. (2004) New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res.*, 32(Web Server issue): W485-91.