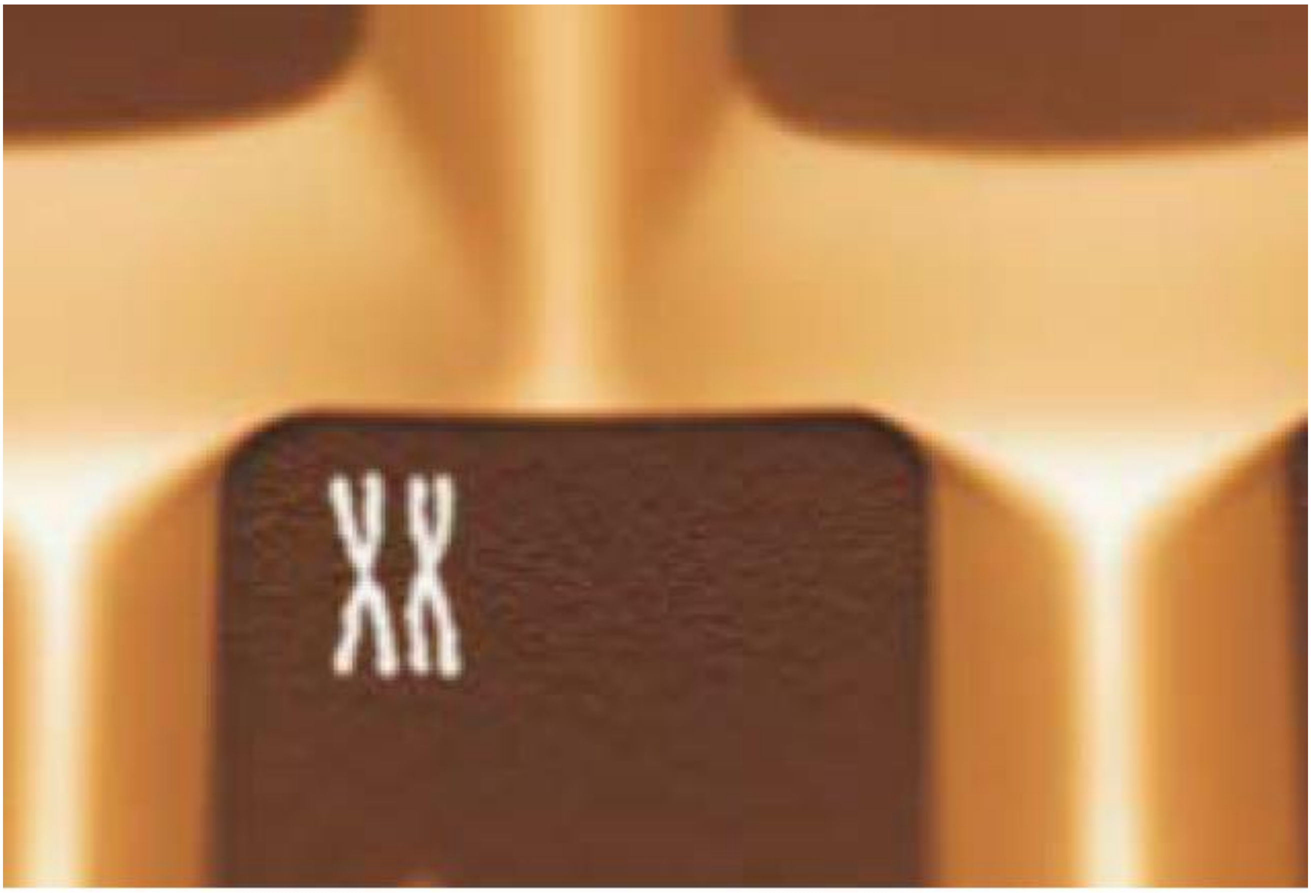

Barcelona, 29 November / 2 December 2004

Proceedings 5th Annual Spanish Bioinformatics Conference

Jornadas de Bioinformática
Jornades de Bioinformàtica
Xornadas de Bioinformatica
Bioinformatikaren jardunaldiak

Editors:
Xavier Messeguer
Gabriel Valiente

UNIVERSITAT POLITÈCNICA DE CATALUNYA
CEPBA
IBM
Genoma España
INB

Xavier Messeguer Gabriel Valiente (Eds.)

5th Annual Spanish Bioinformatics Conference

Jornadas de Bioinformática
Jornades de Bioinformàtica
Xornadas de Bioinformatica
Bioinformatikaren jardunaldiak

Barcelona, November 29–December 2, 2004
Conference Proceedings

Editors

Xavier Messeguer
CEPBA-IBM Research Institute
Technical University of Catalonia
E-08034 Barcelona, Spain
E-mail: peypoch@lsi.upc.es

Gabriel Valiente
Department of Software
Technical University of Catalonia
E-08034 Barcelona, Spain
E-mail: valiente@lsi.upc.es

Proceedings 5th Annual Spanish Bioinformatics Conference

ISBN 84-7653-863-4

Preface

The 5th Annual Spanish Bioinformatics Conference was held in Barcelona (Spain), November 29–December 2, 2004, organized by Feria BioSpain, Technical University of Catalonia, and CEPBA-IBM Research Institute, under the auspicies of Red Temática Nacional de Bioinformática, Instituto Nacional de Bioinformática, Fundación Genoma España, and IBM. Previous editions of the conference were held in Cartagena (2000), Málaga (2001), Salamanca (2002), and A Coruña (2003). The scientific program for all editions of the Annual Spanish Bioinformatics Conference can be found at the respective web pages:

- `http://www.es.embnet.org/~biocomp/bioinfo2000/`
- `http://www.ac.uma.es/~jbi2001/`
- `http://jbi2002.cicancer.org/`
- `http://www.dc.fi.udc.es/ai/otero/jbi03/`
- `http://www.lsi.upc.es/~jbi04/`

The aim of the conference was to provide a common forum between academic and private research institutions to present and discuss their latest results and developments on Bioinformatics and Computational Biology. The scope of the conference covered nearly all aspects of basic and applied research on the field, such as functional and structural genomics and proteomics, comparative genomics and molecular evolution, algorithmic development and computational methods. The conference program included seven sessions of contributed papers, on Functional Analysis (chaired by Joaquín Dopazo), Comparative Genomics (chaired by Andrés Moya), Molecular Evolution (chaired by José Castresana), Structural Analysis and Modeling (chaired by Xavier Avilés), Biomedical Informatics (chaired by Ferran Sanz), Computational Methods (chaired by Gabriel Valiente), and Sequence Analysis (chaired by Roderic Guigó).

The Proceedings of the 5th Annual Spanish Bioinformatics Conference consist of two parts. The first part comprises the 22 contributed papers, out of 45 submissions, that were alloted a 20-minute presentation at the conference. The second part comprises the remaining 23 contributed papers (in submission order) together with the 15 contributed posters (also in submission order) that were presented at the poster session. The acceptance ratio was $22/45 = 49\%$.

We would like to thank the members of the program committee for their help in the selection process. Moreover, we would like to express our gratitude to the local committee members Rosa Badia, Marc Cid, Juan José Porta, and Romà Roset, as well as to the organizing committee members M. Mar Albà, Xavier Avilés, and Julio Rozas.

November 2004

Xavier Messeguer
Gabriel Valiente

Organization

Local Committee

Rosa Badia	AC-CIRI, UPC (Barcelona)
Marc Cid	UPC (Barcelona)
Xavier Messeguer (Chair)	LSI-CIRI, UPC (Barcelona)
Juan José Porta	CIRI, IBM (Barcelona)
Romà Roset	CIRI, UPC (Barcelona)
Gabriel Valiente	LSI, UPC (Barcelona)

Organizing Committee

M. Mar Albà	UPF (Barcelona)
Xavier Avilés	UAB (Barcelona)
Julio Rozas	UB (Barcelona)

Scientific Committee

José Castresana	IBMB, CSIC (Barcelona)
José M. Carazo	PCM, CNB-CSIC (Madrid)
Xavier Daura	IBB, ICREA-UAB (Barcelona)
Xavier de la Cruz	PC-UB (Barcelona)
Javier De las Rivas	CIC, USAL (Salamanca)
Joaquín Dopazo	CNIO (Madrid)
Roderic Guigó	IMIM (Barcelona)
Pedro Larrañaga	UPV-EHU (Donostia)
Andrés Moya	ICBiBE, UV (València)
Baldomero Oliva	GRIB, UPF (Barcelona)
José Oliver	UGR (Granada)
Modesto Orozco	PC, UB (Barcelona)
Angel Ortiz	CBM, UAM (Madrid)
Ramon Otero	(A Coruña)
Albert Sorribas	CMB, UDL (Lleida)
Oswaldo Trelles	AC, UMA (Málaga)
Alfonso Valencia	CNB-CSIC (Madrid)

Sponsoring Institutions

Feria BioSpain, Technical University of Catalonia, CEPBA-IBM Research Institute, Red Temática Nacional de Bioinformática, Instituto Nacional de Bioinformática, Fundación Genoma España, and IBM.

Table of Contents

Functional Analysis

Structural Analysis and Modeling

Comparative Genomics

Molecular Evolution

Biomedical Informatics

Computational Methods

Sequence Analysis

Contributed Posters

UVCLUSTER: Searching For Regularities in Complex Graphs

Vicente Arnau[1], Sergio Mars[2], and Ignacio Marín[2]

[1] Departamento de Informática, Universidad de Valencia, Calle Doctor Moliner, 50, E-46100
Burjassot, Valencia

[2] Departamento de Genética Universidad de Valencia, Calle Doctor Moliner, 50, E-46100
Burjassot, Valencia

Abstract. We describe a new program, called UVCLUSTER, that allows various user-directed analyses of complex undirected graphs. We show how UVCLUSTER can be applied to protein-protein interaction data in order to establish the presence of groups of tightly linked proteins that perform particular biological roles. Applications of UVCLUSTER to other types of data are also discussed.

In the post-genomic era, the extraction of relevant information from massive amounts of biological data has become crucial. For example, there is a great interest in developing classification tools for gene expression data derived from microarrays. Another important set of results refers to protein-protein interaction data, that are rapidly accumulating in several prokaryotic and, especially, eukaryotic species.

Each type of data has its own structure and thus requires its own analytical tools. In the case of protein interaction data, results can be conceptually visualized as graphs in which a large number of elements or nodes, in this case proteins, are connected, if they interact, by edges. By analyzing those undirected graphs, we could obtain several types of significant biological information. For example, we could try to determine when a group of proteins is connected to a degree that is well above average for the whole graph, meaning that they may be performing together a particular cellular function. We could also explore whether different cellular compartments or cellular processes are related or, on the contrary, are totally independent, etc.

However, analyzing graphs in search for regularities is not a trivial matter. The two main problems are:

1) The management of the huge amount of information, with hundreds or even thousands of interconnected proteins, that derives from the analyses of complex genomes. In order to cope with all that information, we need tools to actually establish the graph from all the single, experimentally characterized, protein-protein interactions and also tools for visualizing (at least partially) that graph.

2) The analysis requires the generation of classification tools that allow to determine what kind of structures (e. g. highly-connected set of proteins) are present in the graph.

This second kind of problem is part of a very general question in biology and other sciences, that of how to conveniently classify sets of interrelated items provided some type of distance measure among those items. In the case of protein interaction data, there are no tools to conveniently tackle this problem. Not only the mathematical theory for graphs with small-world properties—such as the ones that many biological entities, and in particular interaction data generate—is still poorly developed but also the typical classification tools used in other fields (numerical taxonomy, phylogenetic analysis, etc) cannot be directly applied to those graphs. The main difficulty to use those methods is the "ties in proximity problem" that appears in datasets in which the distances among many elements are identical. The ties in proximity problem is particularly extreme in protein interaction datasets.

To solve these problems, we have recently generated a new program named UVCLUSTER. This program is designed to allow the exploration of graphs in order to detect clusters of significantly interconnected units. UVCLUSTER uses an iterative hierarchical clustering strategy to establish the strength of the connection between any two elements respect to all the elements of the dataset.

The program has several related functions. First, UVCLUSTER is able to import any file containing protein-protein interactions, even the largest currently available. This file can be then filtered in several ways, allowing the user to select significant elements (e. g. particular groups of proteins). Once those elements have been selected, UVCLUSTER rapidly converts the set of individual interactions in a table of minimal distances among all the elements in the filtered dataset. After that, UVCLUSTER establishes a large number of mathematically equally optimal clustering solutions. Then, the program calculates, with the information from all those solutions, the likelihood for each two units to be grouped together. The program finally generates output files that allow the exploration of the results both directly or by converting the clustering results in dendrograms. All these processes are fast enough as to allow the analysis of sets of up to 1000 proteins on a standard PC computer.

An example of the use of UVCLUSTER is as follows (all results were obtained using a PC with an Intel Pentium IV 2.8 GHz processor and 512 MB RAM memory): A dataset of 4721 proteins and 15210 interactions derived from results obtained in the yeast Saccharomyces cerevisiae was converted into a table of distances by UVCLUSTER in about 14 minutes. This table can be saved for future uses, so this time will not be lost every time that this dataset is analyzed. Then, we selected 376 proteins of that species, known to be involved in 8 modules of gene coexpression and biological function (rRNA processing, ribosomes, proteasome, heat shock, etc). We obtained using UVCLUSTER a total of 10000 alternative clustering solutions for those 376 proteins, in about 5 minutes. The analysis of those solutions rendered 6 main interaction clusters that substantially corresponded to the groups of co-regulated genes acting on particular biological processes previously defined. For a module, the proteasome, for which the set of known protein-protein interactions is large, 29 out of 40 proteins in the module (73%) were also found clustered together using UVCLUSTER.

When applied to protein interaction data, it turns out that UVCLUSTER has four main strengths. First, it allows the characterization of all the sets of closely linked proteins that exist. Second, it can be used to discover the function of unknown proteins, provided we have proteins of known function that interact with them. They all will appear together in statistically significant clusters. Third, it can be used to establish groups of connected proteins even when some information is not available and therefore it can be used to predict potential interactions, to be experimentally tested. Finally, it can be adapted to compare the interactome graphs of different species, considering the orthology relationships among their genes.

UVCLUSTER has been devised for groups interested in functional biology that want to obtain significant information for parts of the interactome of a species in a short time. However, it is obvious that it can be also used to analyze information different from interaction data. The only requirement is for the data to be converted into undirected graphs as the one described above. For example, the connections among protein domains, in which distances are defined according to the combinations of domains present in single proteins in one or multiple species, can be also analyzed using UVCLUSTER. Similarly, our program may be used in non-biological fields, for example to analyze graphs of paper citations or coauthorships.

References

1. Arnau, V., Mars, S., Marin, I.: Iterative cluster analysis of protein interaction data. Bioinformatics (2004) In press.