Introduction

Overview of BITS2005, the Second Annual Meeting of the Italian Bioinformatics Society

Manuela Helmer-Citterich*[1], Rita Casadio[2], Alessandro Guffanti[3], Giancarlo Mauri[4], Luciano Milanesi[5], Graziano Pesole[6], Giorgio Valle[7] and Cecilia Saccone[5,8]

Address: [1]Centre for Molecular Bioinformatics, Dept of Biology, University of Rome Tor Vergata, Rome, Italy, [2]Department of Biology, University of Bologna, Italy, [3]FIRC Institute of Molecular Oncology, Milan, Italy, [4]Department of Computer Science, Systems and Communication, University of Milan-Biococca, Milan, Italy, [5]Institute for Biomedical Technologies, National Research Council (ITB-CNR), Milan, Italy, [6]Dept of Biomolecular Sciences and Biotechnologies, University of Milan, Milan, Italy, [7]Department of Biology and CRIBI Biotechnology Centre, University of Padua, Padua, Italy and [8]University of Bari and ITB - CNR, Bari, Italy

Email: Manuela Helmer-Citterich* - citterich@uniroma2.it; Rita Casadio - citterich@uniroma2.it; Alessandro Guffanti - guffanti@ifom-ieo-campus.it; Giancarlo Mauri - citterich@uniroma2.it; Luciano Milanesi - citterich@uniroma2.it; Graziano Pesole - graziano.pesole@unimi.it; Giorgio Valle - citterich@uniroma2.it; Cecilia Saccone - cecilia.saccone@ba.itb.cnr.it

* Corresponding author

Abstract

The BITS2005 Conference brought together about 200 Italian scientists working in the field of Bioinformatics, students in Biology, Computer Science and Bioinformatics on March 17–19 2005, in Milan. This Editorial provides a brief overview of the Conference topics and introduces the peer-reviewed manuscripts accepted for publication in this Supplement.

Introduction

The ITalian Bioinformatics Society (B.IT.S.; http://bioin formatics.it) aims at bringing together research scientists interested in Bioinformatics, as a multi-disciplinary science for studying biological systems at the molecular and cellular level by using informatics and computational methods and models. The main goals of B.IT.S. are the study, development and spread of Bioinformatics throughout the Italian scientific, academic, technological and industrial community.

The Society was founded in 2003 after 5 years of enjoyable and fruitful informal meetings. It attracts the interest and efforts of scientists from different research areas in Italy: Molecular Biology, Biochemistry, Physics and Computer Science.

The Second Annual BITS Meeting http://www.itb.cnr.it/ bits2005/ was held at the Leonardo Da Vinci Hotel Congress Center in Milan on March 17–19th, 2005. David Lipman, Director of NCBI, Shoshana Wodak, from the Department of Biochemistry, University of Toronto, Canada, and Mikhail Gelfand, from the Institute for Information Transmission Problems, RAS, Russia were the keynote speakers at the BITS2005 Conference. Dr. Lipman opened the meeting with the "Giuliano Preparata Lecture", dedicated to a brilliant and innovative Italian physicist who also made important contributions in the field of Computational Biology, and, unfortunately, passed away a few years ago. Dr. Lipman's talk was about the Semantic Shift in Comparative Genomics & Systems Biology and started an interesting discussion about the new developments and challenges that Bioinformatics will have to face in the near future. Mikhail Gelfand talked about the evolution of riboswitches, while Shoshana

Wodak gave a talk entitled Protein-Protein Interactions: The challenge of prediction specificity, and spoke from a historical perspective with interesting comments about the latest developments in the field.

The conference was organized into thematic sessions: Comparative Genomics, Database and Data Mining, Structural Bioinformatics, Algorithms and Applications, Functional Bioinformatics, Medical Bioinformatics. Abstracts were collected on the different topics and selected for oral presentations; some of them were then submitted as complete research papers for this Supplement. The manuscripts were edited by a committee composed of the Meeting Organizers and the Steering Committee of the Italian Bioinformatics Society and then sent for peer review to a panel of non-Italian referees.

The manuscripts cover several topics central to the development of Computational Biology, such as the development and usage of resources for analysing expression data, tools for the analysis of protein structure, interaction and networks, and the development of informatics resources for biological data integration.

Genome analysis

L. Milanesi *et al.* [1] describe a systematic analysis of the human genome in the search for proteins containing kinase catalytic domains. The predicted set of human kinases, the human *kinome*, was extended by identifying both additional genes and potential splice variants. The results of the research are collected in the KinWeb database, available at the address http://bioinfo.itb.cnr.it/kinweb/.

The ESTree database by Lazzari *et al.* [2], collecting both in house prepared cDNA libraries and publicly available *Prunus persica* expressed sequence tags (ESTs), represents a useful resource for peach functional genomics. The database contains more than 18,000 sequences, with links to predicted SNPs, GO terms, the NiceZyme and KEGG pathway databases.

Sequence analysis

A number of papers describe tools designed for sequence analysis.

ParPEST (Parallel Processing of ESTs) is a pipeline based on parallel computing for EST analysis [3]. It relies on distributed processes and parallelized software.

The GFINDer resource [4] was presented with new modules for performing phenotype analyses of inherited disorder related genes [5]. New GFINDer modules make it possible to annotate large numbers of user classified sequence identifiers with morbidity and clinical information, classifying them according to genetic disease phenotypes and their locations of occurrence, and statistically analyzing the obtained classifications.

Fariselli *et al.* [6] describe the implementation of the posterior-Viterbi (PV) algorithm for the prediction of the topology of all-beta membrane proteins. They show that PV decoding performs better than other algorithms when tested on the problem of predicting the topology of beta-barrel membrane proteins.

A neural network approach is described in Ferraro *et al.* [7] for the inference of SH3 domain specificity. The network performs better than regular expressions and PSSMs in the detection of SH3 domains interactors. The authors show that this approach, however, is dependent upon the number of binding peptides used in the training set.

Phylogenetics

Caenorhabditis elegans and *C. briggsae* share conserved but rapidly evolving genomes. Rambaldi *et al.* [8] developed NemaFootPrinter (Nematode Transcription Factor Scan Through Phylogenetic Footprinting), a web-based software for interactive identification of conserved, non-exonic DNA segments in their genomes. This software relies on the identification of orthologous genes and on the manual selection of gene boundaries. The resource is available at the address: http://bio.ifom-firc.it/NTFootPrinter.

Structural bioinformatics

Secondary structure prediction is considered an interesting topic *per se* and also as an ancillary method to structure prediction methods. Armano *et al.* [9] describe a hybrid technique for secondary structure prediction, based on a "sequence-to-structure" prediction, enforced by resorting to a population of hybrid (genetic-neural) experts, and then on a "structure-to-structure" prediction, by means of an artificial neural network. This new technique attains 76% accuracy, comparable with other state-of-the-art methods.

Secondary structure assignment also represents a fundamental issue, both for classification purposes and for improving the potential of prediction methods. Cubellis *et al.* [10] developed an accurate method for assigning secondary structure based on main chain geometry. The SEGNO program is compared with other pre-existing methods. It defines more types of secondary structure (i.e. poly-proline and 3–10 helices). Moreover, amino-acid trends at helix caps are stronger, secondary structural elements are less likely to be concatenated together, and secondary structure guided sequence alignment is improved.

Ausiello *et al.* [11] describe the Query 3D program for local comparison of protein structures. Query 3D is at the core of the pdbFun server [12] for the identification of local structural similarities between annotated residues in proteins.

D'Ursi *et al.* [13] used a flexible docking approach to characterize the molecular interaction between seven endocrine disrupting chemicals and estrogen, progesterone and androgen receptors in the ligand-binding domain. All ligands docked in the buried hydrophobic cavity corresponding to the hormone steroid pocket. The results are in agreement with known toxicological data and suggest a hydrophobic cavity is needed to accommodate the analyzed chlorine-carrying compounds.

In [14], Greco *et al.* analyze the relatively rare double histone fold, which is tightly related to the structure of nucleosomal histones. Through the application of several secondary structure prediction and fold recognition methods, they showed that the viral protein gi|22788712 is compatible with the structure of a H3-H4-like histone pseudodimer and may retain the ability of mediating protein-DNA interactions.

Gene expression
Several papers in this Supplement concentrate on functional genomics and the analysis of expression data, proving the field's vitality and the experimental laboratory's interest in involving the bioinformatic community in their research.

Ancona *et al.* [15] compare the Regularized Least Squares (RLS) and Support Vector Machines (SVM) approaches in cancer classification through the analysis of microarray data. They show that the results of the two methods are comparable, even if RLS may represent a valuable alternative due to their simplicity and low computational complexity.

Burgarella *et al.* [16] developed MicroGen, a web system for managing information and workflow in the production pipeline of spotted microarray experiments. MicroGen is composed of a core multi-database system able to store all data from different spotted microarray experiments according to the Minimum Information About Microarray Experiments (MIAME) standard. It offers an intuitive and user-friendly web interface able to support collaborative work among the multidisciplinary actors and roles involved in spotted microarray experiment production.

The analysis of two transcription profiles led Cavallo *et al.* [17] to the identification of five Tumor Associated Antigens (TAA), whose expression is linearly related to the tumor mass increase in BALB-neuT mammary glands. Normal expression of these proteins is low and compatible with the design of immunopreventive vaccines.

Finocchiaro *et al.* [18] describe an interesting approach, based on data extracted from biomedical literature and the analysis of Gene Ontology categories. Such data can be used to complement expression data in order to highlight microarray datasets biologically related in gene expression.

Di Camillo *et al.* [19] describe and test a quantization method, based on a model of the experimental error and on a significance level able to mediate between false positive and false negative classifications in the analysis of microarray data. The quantization method, evaluated in comparison with the two standard methods, improves the ability of Reveal and Dynamic Bayesian Networks to identify relations between genes.

Genetic and population analysis
The Hmt database is presented by Attimonelli *et al.* [20], as a well-integrated web-based human mitochondrial bioinformatic resource aimed at supporting population genetics and mitochondrial disease studies. HmtDB consists of a database of Human Mitochondrial Genomes, annotated with population data, and a set of bioinformatic tools. It is able to produce site-specific variability data and to automatically characterize newly-sequenced human mitochondrial genomes.

The PedNavigator tool [21], specifically designed by Mancosu and co-workers for genetic studies, is a browser for genealogical databases. It is useful for genealogical research due to its capacity to represent family relations between individuals and to make a visual verification of the links during family history reconstruction. As for genetic studies, it is helpful to follow propagation of a specific set of genetic markers (haplotype), or to select people for linkage analysis, showing relations between various branches of an affected subjects' family tree.

Systems biology
High-throughput approaches have been applied to the analysis of protein interaction in several model organisms, but have not yet been attempted in humans, where the unraveling of the interactome is one of the most ambitious tasks facing proteomics. An inferred human protein interaction network was built by Persico *et al.* [22], based on the identification of reliable orthologues of proteins known to interact in a number of reference sets. The HomoMINT resulting network is stored in the MINT database [23].

Data and text mining

A high performance workflow is described by Merelli *et al.* [24]. Using grid technology, it correlates different kinds of bioinformatic data, from the nucleotide sequence to the exposed residues of the protein surface. The proposed workflow is implemented to integrate huge amounts of data. The results must be stored in a relational database.

Databases and ontologies

Romano *et al.* [25] describe a system that is able to access and execute predefined workflows. Web Services allow access to the IARC TP53 Mutation Database [26] (containing all TP53 gene mutations identified in human cancers and cell lines that have been reported in the peer-reviewed literature since 1989) and CABRI (Common Access to Biological Resources and Information) catalogues of biological resources.

An SRS site, with both EMBL and CABRI catalogues, has been set up by Romano *et al.* [27]. In the site about 67,500 valid cross-references were identified between the two databases. Such links were added to the EMBL Data Library and now make it possible to establish further links between the CABRI catalogues and other bioinformatic databases cross-referenced in the EMBL database.

Future meeting

The next Annual meeting of the Italian Bioinformatics Society will be held in Bologna in Spring 2006. Further information about BITS2006 will be available on our web site at the address http://bioinformatics.it.

Acknowledgements

References

1. Milanesi L, Petrillo M, Sepe L, Boccia A, D'Agostino N, Passamano M, Di Nardo S, Tasco G, Casadio R, Paolella G: Systematic analysis of human kinase genes: a large number of genes and alternative splicing events result in functional and structural diversity. *BMC Bioinformatics* 6(Suppl 4):S20.
2. Lazzari B, Caprera A, Vecchietti A, Stella A, Milanesi L, Pozzi C: **ESTree db: a Tool for Peach Functional Genomics.** *BMC Bioinformatics* **6(Suppl 4):**S16.
3. D'Agostino N, Aversano M, Chiusano ML: **ParPEST: a pipeline for EST data analysis based on parallel computing.** *BMC Bioinformatics* **6(Suppl 4):**S9.
4. Masseroli M, Galati O, Pinciroli F: **GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists.** *Nucleic Acids Research* 2005, **33:**W717-W723.
5. Masseroli M, Galati O, Manzotti M, Gibert K, Pinciroli F: **Inherited disorder phenotypes: controlled annotation and statistical analysis for knowledge mining from gene lists.** *BMC Bioinformatics* **6(Suppl 4):**S18.
6. Fariselli P, Martelli PL, Casadio R: **A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins.** *BMC Bioinformatics* **6(Suppl 4):**S12.
7. Ferraro E, Via A, Ausiello G, Helmer-Citterich M: **A neural strategy for the inference of SH3 domain-peptide interaction specificity.** *BMC Bioinformatics* **6(Suppl 4):**S13.
8. Rambaldi D, Guffanti A, Morandi P, Cassata G: **NemaFootPrinter: a web based software for the identification of conserved non-coding genome sequence regions between C. elegans and C. briggsae.** *BMC Bioinformatics* **6(Suppl 4):**S22.
9. Armano G, Mancosu G, Milanesi L, Orro A, Saba M, Vargiu E: **A Hybrid Genetic-Neural System for Predicting Protein Secondary Structure.** *BMC Bioinformatics* **6(Suppl 4):**S3.
10. Cubellis MV, Cailliez F, Lovell SC: **Secondary structure assignment that accurately reflects physical and evolutionary characteristics.** *BMC Bioinformatics* **6(Suppl 4):**S8.
11. Ausiello G, Via A, Helmer-Citterich M: **Query 3d: a new method for high-throughput analysis of functional residues in protein structures.** *BMC Bioinformatics* **6(Suppl 4):**S5.
12. Ausiello G, Zanzoni A, Peluso D, Via A, Helmer-Citterich M: **pdb-Fun: Mass selection and fast comparison of annotated PDB residues.** *Nucl Acids Res* 2005, **33:**W133-W137.
13. D'Ursi P, Salvi E, Fossa P, Milanesi L, Rovida E: **Modelling the interaction of steroid receptors with endocrine disrupting chemicals.** *BMC Bioinformatics* **6(Suppl 4):**S11.
14. Greco C, Fantucci P, De Gioia L: **In silico functional characterization of a double histone fold domain from the Heliothis zea virus 1.** *BMC Bioinformatics* **6(Suppl 4):**S15.
15. Ancona N, Maglietta R, D'Addabbo A, Liuni S, Pesole G: **Regularized Least Squares Cancer Classifiers from DNA microarray data.** *BMC Bioinformatics* **6(Suppl 4):**S2.
16. Burgarella B, Cattaneo D, Pinciroli F, Masseroli M: **MicroGen: a MIAME compliant web system for microarray experiment information and workflow management.** *BMC Bioinformatics* **6(Suppl 4):**S6.
17. Cavallo F, Astolfi A, Iezzi M, Cordero F, Lollini PL, Forni G, Calogero R: **An integrated approach of immunogenomics and bioinformatics to identify new Tumor Associated Antigens (TAA) for mammary cancer immunological prevention.** *BMC Bioinformatics* **6(Suppl 4):**S7.
18. Finocchiaro G, Mancuso F, Muller H: **Mining published lists of cancer related microarray experiments: Identification of a gene expression signature having a critical role in cell-cycle control.** *BMC Bioinformatics* **6(Suppl 4):**S14.
19. Di Camillo B, Sanchez-Cabo F, Toffolo G, Nair SK, Trajanoski Z, Cobelli C: **A quantization method based on threshold optimization for microarray short time series.** *BMC Bioinformatics* **6(Suppl 4):**S10.
20. Attimonelli M, Accetturo M, Santamaria M, Lascaro D, Scioscia G, Pappadà G, Russo L, Zanchetta L, Tommaseo-Ponzetta M: **HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research.** *BMC Bioinformatics* **6(Suppl 4):**S4.
21. Mancosu G, Cosso M, Marras F, Cappio Borlino C, Ledda G, Manias T, Adamo M, Serra D, Melis P, Pirastu M: **Browsing Isolated Population Data.** *BMC Bioinformatics* **6(Suppl 4):**S17.
22. Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G: **HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms.** *BMC Bioinformatics* **6(Suppl 4):**S21.
23. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Lett* 2002, **513:**135-140.
24. Merelli I, Morra G, D'Agostino D, Clematis A, Milanesi L: **High performance workflow implementation for protein surface characterization using grid technology.** *BMC Bioinformatics* **6(Suppl 4):**S19.
25. Romano P, Marra D, Milanesi L: **Web services and workflow management for biological resources.** *BMC Bioinformatics* **6(Suppl 4):**S24.
26. Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P: **The IARC TP53 Database: new online mutation analysis and recommendations to users.** *Hum Mutat* 2002, **19:**607-614.
27. Romano P, Dawyndt P, Piersigilli F, Swings J: **Improving interoperability between microbial information and sequence databases.** *BMC Bioinformatics* **6(Suppl 4):**S23.