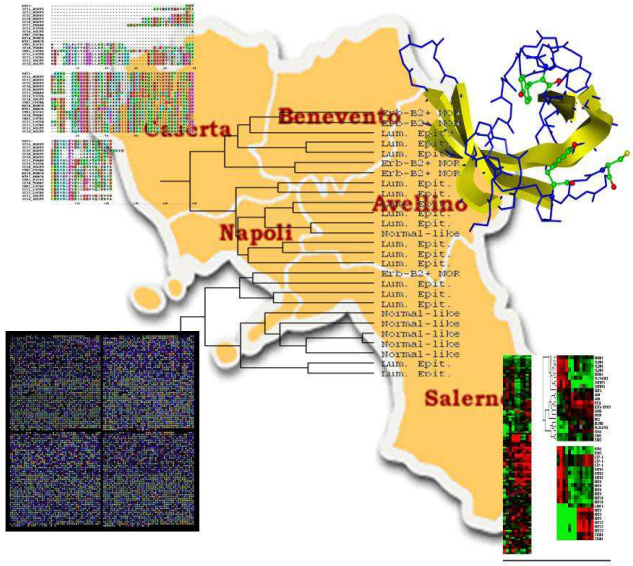
Bioinformatics for Omics Sciences (B4OS)

and

Bioinformatica e Biologia Computazionale in Campania BBCC2012



Area di Ricerca
Consiglio Nazionale delle Ricerche
Napoli
25-27 Settembre 2012

Scientific and Organizing Committee:

- Dott. Angelo Facchiano, Istituto di Scienze dell'Alimentazione, CNR, Avellino
- Dott.ssa Claudia Angelini, Istituto di Applicazioni del Calcolo, CNR, Napoli
- Dott.ssa Maria Luisa Chiusano, Dip. di scienze del suolo, della pianta, dell'ambiente e delle produzioni animali, Università "Federico II", Portici (NA)
- Dott. Mario Guarracino, Istituto di Calcolo e Reti ad Alte Prestazioni, CNR, Napoli
- Dott.ssa Anna Marabotti, IRCCS "E.Medea" Ass. "La Nostra Famiglia", Bosisio Parini (LC) & Istituto di Tecnologie Biomediche, CNR, Segrate (MI)

Under the auspices of:

BITS – Società Italiana di Bioinformatica
http://www.bioinformatics.it

Supported by:

Programma Italia – USA "Farmacogenomica Oncologica"

and

InterOmics – Flagship Project
http://www.interomics.eu/it

http://bioinformatica.isa.cnr.it/BBCC/BBCC2012

Program

25-09-2012

9.00-9.20	**Registration**	
9.20-10.00	*Course Organizers*	**Introduction to the Course**
10.00-11.30	*Participants to the Course*	**2 minutes for short presentation about interests and ongoing projects**
	Pause	
11.45-12.45	**Assunta-Susanna Sansone** University of Oxford	**Data management and curation: the other side of bioinformatics**
	Pause	
14.00-15.30	**Italia De Feis** IAC-CNR, Napoli	**Dimension reduction and classification methods for the analysis of experimental data**
	Pause	
16.00-18.00	**Pedro Jose Navarro Alvarez** Institute of Molecular Systems Biology, Zurich	**Targeted and data independent acquisition approaches in proteomics: computational and statistical challenges to achieve accurate peptide inference**

26-09-2012

9.00-10.45	**Paola Festa** Univ. Federico II, Napoli	**Combinatorial optimization approaches for clustering and biclustering**
	Pause	
11.00-12.00	**Massimo Delledonne** Univ. di Verona	**Comparison of Next Generation Sequencing technologies for genomics and transcriptomics**
12.00-13.00	**Jan Komorowski** Uppsala University	**Monte Carlo feature selection and rough sets. An approach to combinatorial modeling in systems biology**
	Pause	
14.15-15.45	**Alberto Policriti** Univ. di Udine	**Next generation sequencing and methylation profiling**
	Pause	

16.00-18.00	**Paolo Ribeca** Centro Nacional de Análisis Genómico, Barcelona	**Algorithms for high-quality mapping of NGS reads**

27-09-2012

Joint day: **BIOINFORMATICS FOR OMICS SCIENCES Course**
and
BBCC 2012 meeting

9.30-10.00	*Opening*	
10.00-10.30	**Paolo Ribeca** Centro Nacional de Análisis Genómico, Barcelona	**Some take-home messages from genome mappability**
10.30-11.00	**Assunta-Susanna Sansone** University of Oxford	**The reality from the buzz: how to deliver reproducible bioscience data**
	Coffee Break	
11.30-11.50	**Giorgio Giurato** LabMedMolGe Università di Salerno	**iMir: A flexible and automated pipeline for high-throughput miRNA-Seq data analysis**
11.50-12.10	**Francesco Musacchia** Univ. Federico II & Stazione Zoologica "A. Dohrn", Napoli	**Biclustering of gene expression data: metaheuristic algorithms and experimental results**
12.10-12.30	**Most. Mauluda Akhtar** Univ. Federico II, Napoli	**CpG islands under selective pressure are enriched with histone H3 lysine 4 trimethylation**
	Lunch	
14.00-14.20	**Remo Sanges** Stazione Zoologica "A. Dohrn", Napoli	**De-novo generation of the *Octopus vulgaris* reference transcriptome**
14.20-14.40	**Vittorio Fortino** Univ. di Salerno	**Computational methods for predicting transcriptional units in bacteria using different data sources**
14.40-15.00	**Luciana Esposito** IBB-CNR, Napoli	**The classical resonance model does not predict the variability of bond distances and planarity in peptide bonds**
15.00-15.30	Spot presentations (5' each)	*M. Alfieri - R. Esposito - M.R. Coscia - M. Salzano*
15.30	Discussion and conclusions	

B4OS

Abstracts of Lectures

Data management and curation: the other side of bioinformatics

Susanna-Assunta Sansone
University of Oxford, Oxford e-Research Center,
Oxford, UK
http://uk.linkedin.com/in/sasansone

Increased availability of the bioscience data generated is fuelling increased consumption, and a cascade of derived datasets that accelerate the cycle of discovery. But the successful integration of heterogeneous data from multiple providers and scientific domains is already a major challenge within academia and industry. Even when datasets are publicly available, published results are often not reusable due to incomplete description of the experimental details. In the last decade, several data preservation, management, sharing policies, and plans have emerged in response to increased funding for high-throughput approaches in genomics and functional genomics bioscience [1]. A growing number of community-based initiatives have developed minimum reporting guidelines, terminologies and formats (referred to generally as community standards) [2] to structure and curate datasets, enabling data annotation to varying degrees; other efforts work to maximize the interoperability among these standards [e.g. 3, 4]. Researchers and bioinformaticians in both academic and commercial bioscience, along with funding agencies and publishers, embrace the concept that standards are pivotal to enriching the annotation of the entities of interest (e.g., genes, metabolites) and the experimental steps (e.g., provenance of study materials, technology and measurement types), to ensure that shared investigations are comprehensible and (in principle) reproducible.

But despite all these efforts, in practice data sharing is challenging [5]. Vast swathes of bioscience data still remain locked in esoteric formats, are described using ad hoc or proprietary terminology [e.g. 6], or lack sufficient contextual information; many tools do not implement standards — even where these exists; a current wealth of domain-specific reporting standards, or their incompleteness and absence in other areas are other major challenges.

My presentation will provide a snapshot of the current situation. I will highlight a number of stories, the social engineering side and also key

challenges, enriched by my experience over the last decade by working with a variety of stakeholders, including bioscience researchers, bioinformaticians, developers in public and private sectors, standards developing communities, as well as funders and publishers.

References

1. Field D*, Sansone SA*, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, Millard S, Mugabushaka AM, Perrin N, Remacle JE, Remington K, Rocca-Serra P, Taylor CF, Thorley M, Tiwari B, Wilbanks J: Megascience. 'Omics data sharing. Science 326(5950):234-236 (2009)

2. List of standards at BioSharing: http://www.biosharing.org

3. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25(11):1251-1255 (2007)

4. Taylor CF,* Field D*, Sansone SA*, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK Jr, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novère N, et al.: Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26(8):889-896 (2008)

5. Sansone SA and Rocca-Serra P: On the evolving portfolio of community-standards and data sharing policies: turning challenges into new opportunities. GigaScience 1:10 (2012)

6. Harland L, Larminie C, Sansone SA, Popa S, Marshall MS, Braxenthaler M, Cantor M, Filsell W, Forster MJ, Huang E, Matern A, Musen M, Saric J, Slater T, Wilson J, Lynch N, Wise J, Dix I: Empowering industrial research with shared biomedical vocabularies. Drug Discov Today 16(21-22):940-947 (2011)

Dimension reduction and classification methods for the analysis of experimental data

Italia De Feis
Istituto di Applicazioni del Calcolo "M. Picone" (IAC), CNR, Napoli

In the last two decades the high resolution technologies have revolutionized the landscape of molecular biology, demanding for new computational tools to analyze the big quantity of produced data and their relationships. In particular, the hard task of trying to classify the biomolecular profile of some pathologies, cancer, diabetes, etc., to identify new biomarkers have met the so-called "curse of dimensionality", i.e. high dimensional dataset /low dimensional observations. This has implied the development of new strategies to face with this problem.
This lecture will give an overview of the most common classification techniques embedded with dimension reduction tools based on penalty methods. We will discuss the statistical decision theory based on Bayes theorem, the Linear Discriminant Analysis (LDA), the Quadratic Discriminant Analysis (QDA), the Flexible Discriminant Analysis (FDA), the linear Support Vector Machines and the multinomial model together with Lasso-type penalties in order to obtain variable selection. Moreover we will introduce the basic metric measures for classifier performance and the strategies to evaluate it.

Targeted and data independent acquisition approaches in proteomics: computational and statistical challenges to achieve accurate peptide inference

Pedro Navarro
Institute of Molecular Systems Biology, ETH Zürich
navarro@imsb.biol.ethz.ch

Targeted and Data Independent Acquisition mass spectrometry approaches merged as high sensitivity and reproducible technologies addressed to complete a powerful set of tools able to perform deep proteome analysis. SRM (Selected Reaction Monitoring) allows reliable quantification of low abundant peptides in complex samples [1], whereas Data Independent Acquisition provides a higher sample covering and high resolution in time resolved fragment ion monitorization. Despite their potential, both approaches present challenges that have not yet been completely solved.

SRM requires reliable coordinates (monitored fragment ion masses and peptide elution time) for each peptide: obtaining these coordinates can be a tedious and time consuming process. Any targeted (or pseudo-targeted as DIA) analysis requires a thorough protein and peptide selection, observing the uniqueness, and robust occurrence and observability of the peptides [2]. In technological terms, in an SRM experiment there is no precise monitoring of the observed masses, and peptide identification is based on observation of coelution of several child ions of the same parental ion mass in chromatographic profiles. This raises the question: "Can we guarantee that the monitored masses actually come from the targeted parental ion?"

One solution to improve the identification is the use of peptide standard retention times, although they may present a high variability due to the chromatographic system or the sample complexity.

Other solutions are the use of unique ion signatures (UIS) of the monitored peptides [3], or adding equivalent isotopically labeled species (AQUA, DP, SILAC) as a reference, in order to help locating the peak elution of the monitored peptides. Both techniques involve a decrease in sensitivity: UIS does not use the strongest signal transitions, and reference species requires doubling the monitored number of transitions, reducing the acquisition time per transition.

Finally, the increase in complexity of the signals analyzed by these techniques in proteomics makes the use of accurate, automatic peak detection tools [4] critical. This is of particular importance in cases where limited prior information for the targeted peptides is known.

References

1. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol 4 (2008).

2. Reker, D. & Malmstrom, L. Bioinformatic Challenges in Targeted Proteomics. J. Proteome Res. 11, 4393–4402 (2012).

3. Sherman, J., McKay, M. J., Ashman, K. & Molloy, M. P. Unique ion signature mass spectrometry, a deterministic method to assign peptide identity. Mol. Cell Proteomics 8, 2051–2062 (2009).

4. Reiter, L. et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Meth 8, 430–435 (2011).

Combinatorial optimization approaches for clustering and biclustering

Paola Festa
Dipartimento di Matematica e Applicazioni "R. Caccioppoli"
Università degli Studi di Napoli FEDERICO II
Compl. MSA – Via Cintia, Napoli

Clustering algorithms aim to group data such that the most similar objects belong to the same group or cluster, and dissimilar objects are assigned to different clusters. In more detail, they take as input a data set and a similarity (or distance) function over the domain, with the aim of finding a partition of the data into groups of mutually similar elements.
Biclustering is a variant of this task that is needed when the input data comes from two domain sets and some relation over the Cartesian product of these two sets is given. In this case, one could be interested in partitioning each of the sets, such that the subsets from one domain exhibit similar behavior across the subsets of the other domain. Roughly speaking, biclustering can be viewed as simultaneous data clustering and feature selection, i.e., detection of significant clusters and the features that are uniquely associated with them, given that not all features are relevant to certain clusters.
Clustering and biclustering have both generated considerable interest over the past few decades, due to the increasing need of efficiently analyzing high-dimensional gene expression data in several different and heterogenous contexts, such as for example in information retrieval, knowledge discovery, and data mining.
Unfortunately, both the problem of clustering and the problem of locating the most significant bicluster have been shown to be NP-complete. Therefore, given the inner intractability of these problems from a computational point of view, heuristic and metaheuristic approaches are needed to efficiently find good solutions in reasonable running times.
In this talk, combinatorial optimization methods will be presented to approach these problems, including and several efficient metaheuristic algorithms.

Comparison of Next Generation Sequencing technologies for genomics and transcriptomics

Massimo Delledonne
Centro di Genomica Funzionale, Dipartimento di Biotecnologie, Università degli Studi di Verona. Strada le Grazie 15, 37133 Italy

DNA sequencing technology has revolutionized biology and driven a massive acceleration in research and development. The chain termination methodology developed by Fredrick Sanger in the 1970's has, until recently, been the most widely used sequencing method. The demand for genome sequence data using the Sanger methodology drove the formation of sequencing centers and collaboration on a global scale but, due to the scale of the task and the cost, only a limited number of species were covered. Since the late 1990's, researchers in both academia and industry have made efforts to develop alternative approaches for DNA sequence determination, looking to obtain greater sequencing throughput and a cost effective sequencing technology.

The commercial launch of the first massively parallel pyrosequencing platform in 2005 opened the new era of high-throughput genomic analysis, now referred to as next-generation sequencing (NGS). As a massively parallel process, NGS generates hundreds of megabases to hundreds of gigabases of nucleotide sequence output in a single instrument run, depending on the platform. The major commercially available NGS platforms in this new market include Illumina, Roche, Life Technologies, Helicos and Pacific Biosciences. In general, these technologies offer faster and much cheaper (cost-per base / cost-per-read) sequencing than the existing Sanger methodology, but with a shorter read length or higher error rate.

NGS machines are democratizing high throughput sequence generation, enabling investigators to conduct experiments that were previously not technically feasible or affordable. On the one hand this creates great opportunity, sequencing more genomes more quickly, opening up new lines of research and revitalizing others, and potentially deploying sequencing in new technological contexts. On the other hand, the unprecedented volumes of data generate by NGS experiments are posing challenges in terms of data transfer, storage and computational analysis that as a side effect is also leading to advances in bioinformatics and

information technology applied to the solution of biological problems. Nevertheless, NGS technologies are having a striking impact on genomic research and the entire biological field.

Monte Carlo feature selection and rough sets. An approach to combinatorial modeling in systems biology

Jan Komorowski
Program in Computational and Systems Biology
Department of Cell and Molecular Biology, Uppsala University, Sweden
and
Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw, Poland

Machine learning is a well-recognized paradigm in bioinformatics and, more recently, in systems biology. In this talk we will show how the early machine learning is moving from applications to relatively small classifiers and focus on the quality of classification to very large and usually ill-defined systems. In these systems the structure of the classifier, i.e. selection and ranking of significant features together with their combinations and the interpretability of the classifier are sought after by the life scientist. We shall introduce two complementary formalisms Rough Sets of Pawlak and Monte Carlo Feature Selection (jointly with J. Koronacki) and show how they are successfully applied to a wide range of modeling problems in Life Sciences.

Next-Generation Sequencing and methylation profiling

Alberto Policriti
Dip. Matematica e Informatica, Università di Udine

Cytosine methylation is a DNA modification that has great impact on the regulation of gene expression and important implications for the biology and health of several living beings, including humans. Bisulfite conversion followed by next generation sequencing (BS-seq) of DNA is the gold standard technique used to detect DNA methylation at single-base resolution on a genome scale through the identification of 5-methylcytosine (5-mC). However, by converting unmethylated cytosines into thymines, BS-seq poses computational challenges to read alignment and aggravates the issue of multiple hits due to the ambiguity raised by the reduced sequence complexity.
In this talk we will first discuss the basic algorithmic ideas and techniques for sequence alignment, then we will briey introduce the main ideas behind ERNE-BS5 (Extended Randomized Numerical alignEr - BiSulfite 5), an aligning program developed to eficiently map BS-treated reads against large genomes (e.g., human). Three difierent ideas will be illustrated: (i) the use of a 5-letters alphabet for storing methylation information, (ii) the use of a weighted context-aware Hamming distance to identify a T coming from an unmethylated C context, and (iii) the use of an iterative process to position multiple-hit reads starting from a preliminary map built using single-hit alignments. The map is corrected and extended at each cycle using the alignments added in the previous iteration.
ERNE (Extended Randomized Numerical alignEr) is a short string alignment package whose goal is to provide an all-inclusive set of tools to handle short reads. ERNE comprises: ERNE-MAP, ERNE-DMAP, ERNE-FILTER, ERNE-VISUAL, and, from now on, ERNE-BS5. ERNE is free software and distributed with an Open Source License (GPL V3) and can be downloaded at: http://erne.sourceforge.net.

Algorithms for high-quality mapping of NGS reads

Paolo Ribeca
Centro Nacional de Análisis Genómico, Barcelona

Due to the stringent computational requirements, when mapping high-throughput sequencing reads one usually pays more attention to speed than to accuracy. In this lecture we first give a broad overview of the main algorithmic techniques that are currently used for short-read alignment. We then review what are the key points to produce a reliable downstream biological analysis through precise mapping of sequencing data. Finally, we demonstrate how to achieve in practice both high precision and excellent speed with the GEM mapper, a versatile aligner based on FM-indexing and filtration.

BBCC 2012

Abstracts

Some take-home messages from genome mappability

P. Ribeca

Centro Nacional de Análisis Genómico, Barcelona

High-throughput sequencing data is affected by both statistical and systematic errors: among the latter one finds mappability biases. Here we present some general results and caveats obtained by a whole-genome study of mappability in different model organisms.

The reality from the buzz: how to deliver reproducible bioscience data

S.A. Sansone

Principal Investigator, Team Leader
University of Oxford, Oxford e-Research Center, Oxford, UK
uk.linkedin.com/in/sasansone

In this unsettled status quo - presented in my first talk - how can we enable bioscience researchers to make use of existing community standards and maximize data sharing and the subsequent reuse of richly annotated experimental information?
A successful example is provided by the Investigation/Study/Assay (ISA) [1] open source, metadata-tracking framework developed and supported by the growing ISA Commons community [2]. The ISA framework includes both a general-purpose file format and a software suite to tackle the harmonization of the structure of bioscience experimental metadata (e.g., provenance of study materials, technology and measurement types, sample-to-data relationships) by enabling compliance with the community standards. This example illustrates how the synergy between research and service groups in academia, (e.g. in Harvard [3] and at The European Bioinfomatics Institute [4]) and in industry (e.g. at The Novartis Institutes for BioMedical Research and at Janssen Pharmaceuticals, a company of Johnson & Johnson) across a variety of life science domains, is pivotal to build a network of data collection, curation, and sharing solutions that progressively enable the 'invisible use' of standards.
I will present the rationale behind the collaborative development and the evolution of this exemplar ecosystem of data curation and sharing solutions - built on the common ISA framework. I will also provide high-level examples on how this is used to collect, curate and manage heterogeneous experimental metadata in an increasingly diverse set of domains including environmental health, environmental genomics, metabolomics, (meta)genomics, proteomics, stem cell discovery, systems biology, transcriptomics, toxicogenomics, etc.
I will also discuss the experiences learned by my team, our collaborators and the growing user community with usability of the community standards and provide an update on the next steps to develop user-friendly

visualization functionalities and use semantic web approaches to make existing knowledge available for linking, querying, and reasoning.

References

1. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, Neumann S, Sterk P, Tong W, Sansone SA: ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics. 15;26(18):2354-6 (2010); http://isa-tools.org

2. Sansone SA*, Rocca-Serra P*, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L, Begley K, Booth T, Bougueleret L, Burns G, Chapman B, Clark T, Coleman LA, Copeland J, Das S, de Daruvar A, de Matos P, Dix I, Edmunds S, Evelo CT, Forster MJ, Gaudet P, Gilbert J, Goble C, Griffin JL, Jacob D et al.: Toward interoperable bioscience data. Nat Genet 27;44(2):121-126 (2012); http://isacommons.org

3. Ho Sui SJ, Begley K, Reilly D, Chapman B, McGovern R, Rocca-Sera P, Maguire E, Altschuler GM, Hansen TA, Sompallae R, Krivtsov A, Shivdasani RA, Armstrong SA, Culhane AC, Correll M, Sansone SA, Hofmann O, Hide W: The Stem Cell Discovery Engine: an integrated repository and analysis system for cancer stem cell comparisons. Nucleic Acids Res 40(Database issue): D984-91 (2012); http://discovery.hsci.harvard.edu

4. Haug K; Salek R; Conesa P, Hasting J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P, Maguire E, Gonzalez Beltran A, Sansone SA, Griffin J, Steinbeck C: MetaboLights – An open-access general-purpose repository for Metabolomics studies and associated meta-data. Nucleic Acids Res (in review); http://www.ebi.ac.uk/metabolights

iMir: A flexible and automated pipeline for high-throughput miRNA-Seq data analysis

G. Giurato[1], M.R. De Filippo[2], A. Rinaldi[1], C. Cantarella[1], G. Nassa[1], M. Ravo[1], F. Rizzo[1], R. Tarallo[1], A. Weisz[1,3]

[1]*Laboratory of Molecular Medicine and Genomics, Department of Medicine and Surgery, University of Salerno, Baronissi, Salerno, Italy.*
[2]*Fondazione IRCCS SDN, Napoli, Italy.*
[3] *Division of Molecular Pathology and Medical Genomics, "SS. Giovanni di Dio e Ruggi d'Aragona" Hospital, University of Salerno, Salerno, Italy.*

miRNA-Seq represents a novel technology widely used to investigate with high sensitivity and specificity small non-coding RNA populations, comprising microRNAs and other regulatory transcripts. To gather biologically relevant information, such as detection and expression analysis of known and new small non-coding RNA, identification of variants and prediction of their putative targets, the analysis of miRNA-Seq data requires the implementation of multiple statistical and bioinformatics tools from different sources, each focusing on a single step of the analysis pipeline. As consequence, the analytical workflow is slowed down by the need for continuous interventions by the operator, a critical factor when a large number of samples needs to be analyzed at once.

We devised a way to overcome this problem by designing a novel modular pipeline, called iMir, for comprehensive analysis of miRNA-Seq data, from linker removal and sequence quality checks to differential expression and biological target prediction, integrating multiple open source modules and resources linked together in an automated flow. iMir proved to be more efficient and time-effective than most currently available methods and, in addition, it is flexible enough to allow the operator to select the preferred combination of analytical steps. The pipeline was applied to analyze at once 6 miRNA-Seq datasets of 5-7 million tags/run, obtained from exponentially growing or growth-arrested human breast cancer MCF-7 cells, leading to the rapid and accurate identification of > 350 expressed microRNAs, including several differentially expressed in the two conditions tested and their putative mRNA targets, as well as several putative novel microRNAs and isomiR.

iMir is reliable, flexible and fully automated workflow useful to analyze rapidly and efficiently high throughput miRNA-Seq data, such as produced by the most recent high-performance next generation sequencers.

Research supported by: AIRC (Grant IG-8586), MIUR (PRIN 2008CJ4SYWfi004), Regione Campania, University of Salerno (FARB 2011), Fondazione con il Sud, EU COST (Action BM1006 'SeqAhead'), Fondazione Veronesi. Giorgio Giurato is a student of PhD School in Experimental and Clinic Medicine / Doctorate in Experimental Physiopathology and Neuroscience, Second University of Naples (Italy).

Biclustering of gene expression data: metaheuristic algorithms and experimental results

F. Musacchia[1], A. Marabotti[2], A. Facchiano[3], L. Milanesi[2], P. Festa[1]

[1]*Dipartimento di Matematica e Applicazioni "R. Caccioppoli", Università degli Studi di Napoli FEDERICO II, Italy*
[2]*Istituto di Tecnologie Biomediche – CNR, Segrate (MI), Italy*
[3]*Istituto di Scienze dell' Alimentazione - CNR, Avellino, Italy.*

Given a gene expression data matrix, a bicluster is a submatrix of genes and conditions where these have a correlation of expression activity across rows or columns.
This topic had generated considerable interest over the last years, particularly related to the analysis of high-dimensional gene expression data in information retrieval, knowledge discovery, and data mining [1].
The identification of a correct bicluster has been shown to be an NP-complete problem. To solve this problem, we have designed several hybrid metaheuristic approaches, including GRASP-like algorithms and Iterated Local Search.
A GRASP (Greedy Randomized Adaptive Search Procedure) [2,3,4] is a randomized multistart iterative metaheuristic consisting of two phases: a construction phase and a local search phase. The construction phase builds iteratively a feasible solution in a greedy randomized way. Once obtained a feasible solution, a local search procedure tries to improve it by taking the best candidates locally. The two phases, construction and local search, are repeated and the best local optimum is returned as final solution.
We implemented several GRASP-like algorithms that differ in both construction and local search phase. In the construction phase, one implements a k-means and a second one a greedy randomized procedure based on a minimum spanning tree of a suitable weighted graph. Then, two types of local searches have been implemented: one has been already proposed [5] and a second one is an Iterated Local Search [6]. All the designed algorithms have been tested and compared using the Lymphoma [7], Yeast [8], and Arabidopsis [9] datasets.
We have tested our proposals on several different datasets using the mean squared residue [1] as measure of evaluation. The results with Lymphoma, Yeast, and Arabidopsis are on the whole encouraging also from a

biological point of view and pushing the authors to pursue the metaheuristics way.

A.M. and L.M. are supported by MIUR FIRB ITALBIONET (RBPR05ZK2Z and RBIN064YATfi003). The work has been made in the frame of the Flagship Project InterOmics.

References

[1] Y.Cheng and G. Church. (2000) Biclustering of Expression Data, Proc. Int. Conf, Intell. Syst. Mod. Biol., 93-103.

[2] T.A. Feo and M. G.C. Resende (1995) Greedy Randomized Adaptive Search Procedures, J. Global Optim., 6, 109–134.

[3] P. Festa and M.G.C. Resende (2009) An annotated bibliography of GRASP - Part I: algorithms, International Transactions in Operational Research, 16, No. 1, pp. 1-24, 2009.

[4] P. Festa and M.G.C. Resende (2009) An annotated bibliography of GRASP - Part II: applications, International Transactions in Operational Research, 16, No. 2, pp. 131-172, 2009.

[5] F. Musacchia, A. Marabotti, A. Facchiano, L. Milanesi, and P. Festa (2011) Biclustering of gene expression data based on GRASP-like algorithms. BITS2011, ISBN 978-884673069-5, pp. 100-101.

[6] W.Ayadi, M. Elloumi and J.-K. Hao (2010) Iterated Local Search for Biclustering of Microarray Data. Lect. Notes Comput. Sci., 6282, 219-229.

[7] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Jr Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, L.M. Staudt (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403, 503-511.

[8] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, P. O. Brown, Genomic expression programs in the response of yeast cells to environmental changes. Molecular Biology of Cell, vol.11, pp. 4241-4257, 2000.

[9] S. Tavazoie, J.D. Hughes, M. J. Campbell, R. J. Cho, G. M. Church, Systematic determination of genetic network architecture. Nat. Genet., vol.22, pp. 281–285, 1999.

CpG islands under selective pressure are enriched with histone H3 lysine 4 trimethylation

M.M. Akhtar[1,4], S. Cocozza[1,4], G. Miele[1,2,3], A. Monticelli[5]

[1]*Gruppo Interdipartimentale di Bioinformatica e Biologia Computazionale, Università di Napoli "Federico II" - Università di Salerno, Naples, Italy*
[2]*Dipartimento di Scienze Fisiche, Università degli Studi di Napoli "Federico II", Naples, Italy*
[3]*Istituto Nazionale di Fisica Nucleare – Sezione di Napoli, Naples, Italy*
[4]*Dipartimento di Biologia e Patologia Cellulare e Molecolare "L. Califano", Università degli Studi di Napoli "Federico II", Naples, Italy*
[5]*Istituto di Endocrinologia ed Oncologia Sperimentale, CNR Napoli, Naples, Italy*

Histone modification is an epigenetic mechanism that influences gene regulation in eukaryotes. In particular, the enrichment of histone H3 lysine 4 trimethylation (H3K4me3) in CpG islands (CGIs) is known to be associated with the open chromatin state and with transcription activity. Changes in gene expression play a crucial role in adaptation and evolution. In this paper, we have studied, using a computational biology approach, the relation between H3K4me3 enrichment in CGIs and signatures of selective pressure in Homo sapiens. To evaluate H3K4me3 in CGIs, we used the publicly available genomic-scale analyses of histone modifications in three human cell lines. To define regions under selective pressure, we used three distinct approaches that mark the selective events that occurred in three different evolutionary periods. We found that, regardless of the chosen approaches and cell types used, CGIs under selective pressure showed significant enrichment in H3K4me3. As a further check we performed the same analysis on H3K36me3 obtaining a partial support to our finding. Several studies have reported an inverse correlation of H3K4me3 with DNA methylation and the finding reported here supports a previous study in which we found that CGIs under selective pressure were less methylated.

De-novo generation of the *Octopus vulgaris* reference transcriptome

Swaraj Basu, Giuseppe Petrosino, Giovanna Ponte, Ilaria Zarrella, Raffaele Calogero, Graziano Fiorito, Remo Sanges

Stazione Zoologica Anton Dohrn, Villa Comunale – 80121, Naples - Italy

Cephalopod mollusks present the most complex nervous systems outside the vertebrate lineage, therefore it has been often suggested that their genome and transcriptome sequences will provide useful insights into the evolution of complex brains. Recently, the number of noncoding genomic elements has been related to the complexity of the nervous system. Indeed, although the number of coding elements remains relatively stable across all metazoan species, the number of noncoding sequences significantly increases in relation to the complexity. Such elements have been shown to be enriched in proximity to brain expressed genes in mammals and it would be extremely interesting to isolate and analyze them in cephalopods. Therefore, we have started to generate the reference transcriptome of the *Octopus vulgaris* and are planning to sequence its genome. Initial results show intriguing molecular adaptations of different tissues which can explain why, among cephalopods, the octopus is doubtless the most intelligent.

Computational methods for predicting transcriptional units in bacteria using different data sources

V. Fortino[1,2], R. Tagliaferri[1]

[1]Dipartimento di Informatica e [2]Dipartimento di Scienze Farmaceutiche e Biomediche, Università di Salerno

According to the standard definition, operons can be defined as a set of consecutive genes on the same transcriptional strand of a genome sequence, where the genes are co-transcribed into one single mRNA molecule coding for more than one protein.
The identification of genes that are grouped together into operons is a key step to elucidate gene regulation in bacterial genomes. However, the mechanisms of operon formation are poorly understood and experimental methods to identify operon structures are very difficult to implement (Walters et al., 2001). For this reason developing computational methods to effectively predict operons has become a very important challenge in computational biology. Operon prediction is essential not only because it provides the prediction about which genes are co-regulated, but also because the prediction of other regulatory elements, such as transcription binding sites, promoters, etc., often relies on the delineation of operon structures. Besides, operon prediction can improve computer annotation of genomes and helps to infer the function for uncharacterized proteins.
Most of the current prediction methods uses static sources of information, such as intergenic distance (Salgado et al., 2000; Moreno-Hagelsieb and Collado-Videset al., 2002), conserved gene clusters (Overbeek et al., 1999; Tamames et al., 1997), functional relations (Taboada et al., 2010), gene microarray expression data (Sabatti et al., 2002), and combination of several genomic properties (Dam et al., 2006). Unfortunately, recent whole-transcriptome RNA-seq analysis of prokaryotic organisms reveals that using only genomic properties is not sufficient to have accurate operon predictions (Kumar et al., 2012; Hövik et al., 2012), and that, in order to improve classification results, dynamic data sources are necessary. Here we present a novel method that combines transcript borders, obtained from mapped RNA-Seq reads, and standard operon predictions for the identification of all operons revealed in a whole transcriptome profile. So

far the experimental results indicate that we achieve an accurate operon maps including highly reliable predictions of new operon pairs.

References

Dam, P., Olman, V., Harris, K., Su, Z. and Xu, Y. (2006) Operon prediction using both genome-specific and general genomic information. Nucleic Acids Res., 35, 288–298.

Hövik, H., Yu, W.H., Olsen, I., Chen, T. (2012) Comprehensive transcriptome analysis of the periodontopathogenic bacterium Porphyromonas gingivalis W83, J Bacteriol, 194, 100–114.

Kumar, R., Lawrence, M.L., Watt, J., Cooksey, A.M. and Burgess, S.C. (2012) RNA-Seq Based Transcriptional Map of Bovine Respiratory Disease Pathogen Histophilus somni 2336. PLoS ONE, 7, e29435.

Moreno-Hagelsieb, G. and Collado-Vides, J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics, 18, 329–36.

Overbeek, R., Fonstein, M., DSouza, M., Pusch, G.D. and Maltsev, N. (1999) The use of gene clusters to infer functional coupling, Natl Acad. Sci. USA. 96, 2896–2901.

Sabatti, C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res, 30, 2886–2893.

Salgado, H. (2000) Operons in Escherichia coli: genomic analyses and predictions. Proc Natl Acad Sci USA, 97, 6652–6657.

Walters, D.M. (2001) High-density sampling of a bacterial operon using mRNA differential display. Gene, 273, 305–315.

The classical resonance model does not predict the variability of bond distances and planarity in peptide bonds

R. Improta, L. Vitagliano, L. Esposito

Institute of Biostructures and Bioimaging, CNR, Napoli– Italy

Understanding the physico-chemical principles underlying protein structures is a major issue in structural biology. Protein folded states often combine structural complexity with an intrinsic fragility, which is essential for functionality and turnover. Therefore, a full comprehension of the principles that dictate protein structures cannot prescind from the quantification/definition of all factors involved. In this scenario,
The elucidation of peptide bond geometrical properties has proven to be crucial for the prediction of basic elements of protein structure. Indeed, the seminal work by Pauling on peptide bond planarity developed by the application of the resonance model has been crucial for predicting the structure of protein secondary structure elements. More recently, the analysis of high resolution protein structures have highlighted an intricate picture of the interplay of peptide bond geometrical parameters [1-5].
By combining quantum-mechanical analysis on very small model compounds and statistical surveys of protein/peptide structure database we have recently unveiled the stereoelectronic effects associated with peptide group distortions in peptides and proteins [6]. Here, these analyses have been extended to more subtle structural features as bond length and bond angles variability detected in peptide bonds. The excellent agreement between computed and statistical data suggests that peptide bond variability is essentially driven by local effects. Moreover, the analysis of the variability of bond distances and planarity in peptide bonds shows that simple interpretative models based on ‘Lewis like’ pictures, as those used in the classical resonance notation, cannot give full account of subtle structural details observed in real protein structures.

References

1. L. Esposito, L. Vitagliano, A. Zagari, L. Mazzarella. Protein Sci. (2000) 9, 2038.

2. L. Esposito, L. Vitagliano, A. Zagari, L. Mazzarella. Protein Eng. (2000) 13, 825.
3. L. Esposito, A. De Simone, A. Zagari, L. Vitagliano. J. Mol. Biol. (2005) 347, 483.
4. D.S Berkholz, M.V. Shapovalov, R.L. Dunbrack, P.A. Karplus. Structure (2009) 17, 1316.
5. D.S. Berkholz, C.M. Driggersa, M.V. Shapovalovc, R.L. Dunbrack, P.A. Karplus. PNAS (2012)109, 449.
6. R. Improta, L. Vitagliano, L. Esposito. Plos One. (2011) 6, e24533.

RNA-sequencing as a tool for the identification of candidate genes involved in diterpenes biosynthesis in hairy roots of *Salvia sclarea*

M.E. Alfieri, M.C. Vaccaro, V.E. Ocampo, A. Leone

Planta*LAB, Department of Pharmaceutical and Biomedical Sciences, Università degli Studi di Salerno, Fisciano (SA)*

The roots of *S. sclarea* are rich in abietane diterpenoids (*e.g.* aethiopinone, 1-oxoaethiopinone, salvipisone, and ferruginol), with known antibacterial, antifungal, and sedative pharmacological properties. More recently, this class of tricyclic diterpenoids has raised much attention for its cytotoxic activity against human leukemic cell lines (Rozalski et al. 2006 *Z. Naturforsch*., 61, 483-488). Our preliminary results have indicated a cytotoxic effect also on different solid tumor cell lines. These data have prompted us to enhance the biosynthesis of this class of compounds by metabolic engineering. The elucidation of genes involved in the diterpenoid biosynthesis pathway represents the first step to improve the production of these compound. However, the genome of *Salvia sclarea* as well as molecular mechanisms underlying the secondary metabolism in this plant are still largely unknown. To address this goal we aim at analyzing the transcriptome of *Salvia sclarea* by RNA-seq technology. On the basis of our previous results showing increased diterpenoid production after elicitation with methyl-jasmonate (MeJa), we plan to compare the expression profile of MeJa treated- to untreated hairy roots to discover candidate genes involved in the regulation of the abietanic diterpene pathway and targeting endogenous genes to optimize the synthesis of these promising antitumoral molecules.

ALE-HSA21: a pilot integrated and dynamic web portal for human chromosome 21

R. Esposito[1], D. Evangelista[2], M. Scarpato[1], M.R. Ambrosio[1], M. Aprile[1], R. Aversa[1], C. Angelini[2], A. Ciccodicola[1], V. Costa[1].

[1]*CNR, Institute of Genetics and Biophysics "A. Buzzati-Traverso" (IGB), Naples, Italy.*
[2]*CNR, Istituto Per le Applicazioni del Calcolo "Mauro Picone" (IAC), Naples, Italy.*

Transcriptome studies by RNA-Sequencing have disclosed new classes of RNAs, also indicating most of transcription occurs outside gene boundaries. It suggests that correct gene annotations are crucial to bridge the gap between sequence and biology. These considerations are relevant particularly in genetics studies, for complex diseases or genetic syndromes whose etiology is mostly unknown.
In the last years, our research group has focused on the trisomy of human chromosome 21 (HSA21), the chromosomal basis of Down syndrome (DS). Our recent pilot RNA-Seq experiment on trisomic cells has highlighted the importance of novel genetic elements, both coding (new DS-specific splice isoforms) and non-coding RNAs (miRNAs, lincRNAs), and of 5' and 3' untranslated regions (UTRs) of coding genes. Therefore, starting from sequencing data, we focused on HSA21 to identify peaks of "high-density" reads coverage in intronic and intergenic regions, unannotated extended 5' and 3' UTRs and novel splice isoforms, whose validation is still ongoing.
However, the huge amount of data provided by large-scale studies, the lack of a unique user-friendly web resource merging gene annotations and functional data, make difficult to access complete information about one or few genes without browsing different databases.
For these reasons, we are implementing a web resource with a user-friendly interface - for biologically- and computationally-oriented researchers as well as physicians - consisting in a quick retrieval resource for comprehensive and updated analysis of several gene features.
We firstly developed an integrated database, ALE-HSA21 (AnaLysis of Expression of HSA21), divided in 5 main sections, whose core is represented by "Gene" section, consisting of: i) a detailed gene description, including (for each splice variant) reference number, genomic coordinates,

information about encoded protein and the involvement in human genetic diseases; ii) dynamic graphical representations of genes' structures; iii) nucleotide sequences of coding exons, introns, 5'/3' UTRs and promoters, easily downloadable for each transcript, including novel splice isoforms, extended gene boundaries for some HSA21 genes identified by our RNA-Seq study and validated by RT-PCR (in progress); iv) a systematic *in silico* characterization of transcription factors' binding sites in gene promoters, of exonic and intronic regulatory elements and of miRNAs' regulatory binding sites in 3' UTRs. Moreover, links to biological databases such as Genome Browser, dbSNP, OMIM and Gene expression Atlas, are also available, to provide a wider overview on proposed contents.

Secondly, the "Analysis" section, directly linked to the web portal core, allows searching and filtering for each of the above-described features in order to permit a quicker and easier consultation, by selecting only data of users' interest.

This web resource will be a useful support to help scientists in the investigation of HSA21 content with a clear - but not limited to - focus on Down syndrome studies. Such pilot web portal will represent a good model for further analyses aimed to collect and share large-scale gene annotation studies. Finally, further annotation data will be soon produced by RNA-Seq experiments, providing new information to update our web portal.

Molecular structure and dynamics of the complement component C3 of Antarctic teleosts

M.R. Coscia[1], S. Varriale[1], D. Melillo[2], U. Oreste[1], M.R. Pinto[2]

[1] *Institute of Protein Biochemistry, CNR Napoli*
[2] *Stazione Zoologica Anton Dohrn, Napoli*

Complement System is part of the innate immune system; its major function is recognition and elimination of pathogens via direct killing and/or stimulation of phagocytosis. The key molecule of the system is the third component C3. Its activation results in the production of two proteolytic fragments: a small fragment, the anaphylatoxin C3a, that partecipates in several immunological activities, and a large fragment, C3b; the latter undergoes a significant conformational change which leads to the exposure of the active site that comprises a thioester bond whose hydrolysis is catalized by a histidine residue spatially proximal.
We were aimed at extending the knowledge of C3 to two Antarctic teleosts, *Trematomus bernacchii* and *Chionodraco hamatus*, which have been chosen as model species to study the cold-adaptation of the immune system. Two *T. bernacchii* C3-like clones, TbC3-1 and TbC3-2, and one *C. hamatus* clone, ChC3, have been isolated and sequenced. The deduced amino acid sequences showed all the features of the mammalian counterpart, however TbC3-2 lacked the catalytic histidine.
Molecular models of TbC3-1 and TbC3-2 as well as of the C3 molecule from the temperate species *Paralicthys olivaceus* have been built using the crystallographic structure 2A73 of human C3 as a template. *T. bernacchii* C3-1 and *P. olivaceus* C3 models were validated using the PROCHECK programme and minimized with the GROMACS3.2 package. Models of *T. bernacchii* C3a and C3b have also been built. An analysis of the active sites of both C3b models suggested that also TbC3-2, although lacking the histidine residue, could be functional. Molecular Dynamics (MD) simulations have been performed using the GROMACS96 force field and the flexibility of the Cα atoms of the backbone has been compared at the equilibrium. Significant differences in the RMSF plots have been observed. MD simulations of TbC3-1a showed high flexibility in the C-terminal helix, which is the site interacting with the C3a receptor. This

may be interpreted as the result of the dynamic evolutionary process leading to cold-adaptation.

*Work supported by PNRA funding.

The role of aryl hydrocarbon receptor as a chemosensor molecule

M. Salzano[1], A. Marabotti[2], L. Milanesi[2], A. Facchiano[1]

[1]*Istituto di Scienze dell' Alimentazione - CNR, Avellino, Italy*
[2]*Istituto di Tecnologie Biomediche – CNR, Segrate (MI), Italy*

Humans are constantly exposed to an enormous number of chemical molecules present in their environment, that enter the cell and can affect cellular function by either non-selective binding to cellular macromolecules or by interference with cellular receptors. One of these intracellular chemosensor molecules is the aryl hydrocarbon receptor (AhR), a transcription factor of the basic helix-loop-helix / Per-ARNT-Sim (bHLH/PAS) family that is known to mediate the biochemical and toxic effects of dioxins, polyaromatic hydrocarbons and related compounds.
We applied computational methods to simulate the structure of human AhR-Ligand Binding Domain (hAhR-LBD), including PASB and PAC regions, and to characterize interaction of this protein domain with different ligands. The model of hAhR-LBD obtained by homology modelling was used for docking simulations with some among the most important and potent AhR ligands. For each molecule tested, a specific "*binding fingerprint*" was traced. These data allowed to identify the most important residues for the binding of xenobiotics in the hAhR-LBD domain.

References:

Salzano M, Marabotti A, Milanesi L, Facchiano A. (2011) Biochem Biophys Res Commun. 413, 176-181.