CURRICULUM VITAE

General Information

First Name	**Tiziana**
Last Name	**Sanavia**
Citizenship	Italian
Date and place of birth	26/04/1984, Dolo (Venice), Italy

Current Position

Dates Attended	January 2014 - December 2015
Position	**Senior Post-doctoral Fellow**
Scientific field	**Bioengineering**
Institution	Department of Information Engineering, University of Padova (Italy)
Research Activity	Integrative epigenomic and genomic computational methods for high-throughput sequencing data
Supervisor	Ph.D. Barbara Di Camillo

Education

Dates Attended	January 2009 - December 2011
School	**Ph.D. School on Information Engineering (Bioengineering)**
Institution	Department of Information Engineering, University of Padova (Italy)
Thesis	Biomarker lists stability in genomic studies: analysis and improvement by prior biological knowledge integration into the learning process
Supervisor	Ph.D. Barbara Di Camillo
Dissertation	March, 23th 2012
Date	July, 8th 2008
Degree	**Master Degree**
Description	**Bioengineering**
Institution	University of Padova (Italy)
Thesis	Function-based discovery of temporal gene expression patterns in endothelial cells stimulated with insulin.
Advisor	Ph.D. Barbara Di Camillo
Graduation	110/110 cum laude
Date	July, 27th 2006
Degree	**Bachelor Degree**
Description	**Information technologies engineering**
Institution	University of Padova (Italy)
Thesis	Compartmental models: identification and parameter estimation
Advisors	Prof. Claudio Cobelli (advisor) and Ph.D. Chiara Dalla Man (co-advisor)
Graduation	105/110

Research Experiences

Period	January 2012 - December 2013
Activity	**Junior Post-doctoral Fellow** for research projects related to the development of computational methods for the integration of biological information in un-supervised and supervised classification of high-throughput genomic data.
Institution	Department of Information Engineering, University of Padova (Italy)
Supervisor	Ph.D. Barbara Di Camillo

Period	April 2012 - March 2013
Activity	**Visiting scientist** for research projects related to genomic analysis and computational methods for studying the development of pancreatic cells using NGS (Next Generation Sequencing) data: RNA and methylation sequencing.
Institution	Computational Biology and Informatics Laboratory (CBIL), Perelman School of Medicine at University of Pennsylvania (Philadelphia, USA)
Supervisor	Prof. Christian J. Stoeckert Jr.

Period	October 2008 – December 2008
Activity	Scholarship on research project **"Systems Biology approaches to infer gene gene regulation from gene and protein time-series expression data"**
Institution	Department of Information Engineering, University of Padova (Italy)
Supervisor	Ph.D. Barbara Di Camillo

Publications

Peer-reviewed journal papers

[J1] SANAVIA T, FINOTELLO F, DI CAMILLO B (2015) **FunPat: function-based pattern analysis on RNA-seq time series data.** BMC Genomics, 16 Suppl 6:S2.

[J2] SINIGAGLIA A, LAVEZZO E, TREVISAN M, SANAVIA T, DI CAMILLO B, PETA E, SCARPA M, CASTAGLIUOLO I, GUIDO M, SARCOGNATO S, CAPPELLESSO R, FASSINA A, CARDIN R, FARINATI F, PALU' G, BARZON L (2015) **Changes in microRNA expression during disease progression in patients with chronic viral hepatitis.** Liver Int., 35(4):1324-33.

[J3] ZYCINSKI G, BARLA A, SQUILLARIO M, SANAVIA T, DI CAMILLO B, VERRI A (2013) **Knowledge Driven Variable Selection (KDVS) - a new approach to enrichment analysis of gene signatures obtained from high-throughput data.** Source Code for Biology and Medicine 8:2.

[J4] AGHAEEPOUR N, FINAK G, the FlowCAP Consortium, the DREAM Consortium (SANAVIA T within the DREAM consortium), HOOS H, MOSMANN TR, GOTTARDO R, BRINKMAN RR, SCHEUERMANN RH (2013). **Critical Assessment of Automated Flow-Cytometry Analysis Techniques.** Nature Methods, 10(3):228-38.

[J5] SANAVIA T, AIOLLI F, DA SAN MARTINO G, BISOGNIN A, DI CAMILLO B (2012) **Improving biomarker list stability by integration of biological knowledge in the learning process.** BMC Bioinformatics 13 Suppl 4:S22.

[J6] DI CAMILLO B, SANAVIA T, MARTINI M, JURMAN G, SAMBO F, BARLA A, SQUILLARIO M, FURLANELLO C, TOFFOLO G, COBELLI C (2012) **Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment.** PLoS ONE 7(3):e32200.

[J7] DI CAMILLO B, IRVING BA, SCHIMKE J, SANAVIA T, TOFFOLO G, COBELLI C, NAIR KS (2012) **Function-based discovery of significant transcriptional temporal patterns in insulin stimulated muscle cells.** PLoS ONE 7(3):e32391.

[J8] DI CAMILLO B, SANAVIA T, IORI E, BRONTE E, RONCAGLIA E, MARAN A, AVOGARO A, TOFFOLO G, COBELLI C (2010) **The Transcriptional Response in Human Umbilical Vein Endothelial Cells Exposed to Insulin: a Dynamic Gene Expression Approach.** PLoS ONE 5(12):e14390.

Book chapters

[B1] SAMBO F, SANAVIA T, DI CAMILLO B (2013) **Integration of Genetic Variation as External Perturbation to Reverse Engineer Regulatory Networks from Gene Expression Data.** In: De La Fuente A (ed). Gene Network Inference. Springer Berlin Heidelberg, pp. 107-118.

Conference Proceedings

International conferences

[C1] SANAVIA T, FINOTELLO F, DI CAMILLO B (2014) **FunPat: a function-based pattern analysis pipeline for RNA-seq time-series data.** In: BITS Annual Meeting 2014, February 26th-28th 2014, Rome, Italy. ORAL PRESENTATION

[C2] SANAVIA T, MINA M, DI CAMILLO B, GUERRA C, TOFFOLO G (2012) **Recovering stable biomarker lists using a network-based measure of connectivity from Protein-Protein Interactions.** In: Intelligent Data Analysis in Biomedicine And Pharmacology (IDAMAP). November 22nd 2012, Pavia - Italy.

[C3] MINA M, SANAVIA T (2012) **FastSemSim: Fast SEMantic SIMilarity over Gene Ontology annotations.** In: 10th European Conference on Computational Biology (ECCB). September 9th-12th 2012, Basil - Switzerland.

[C4] FONTANA P, SANAVIA T, FACCHINETTI A, LAVEZZO E, FALDA M, CAVALIERI D, DI CAMILLO B, TOPPO S (2012) **Can we go beyond sequence similarity to predict protein function?** In: Automated Function Prediction SIG 2012: Critical Assessment of Function Annotations (CAFA). July 14th 2012, Long Beach - California (USA).

[C5] ZYCINSKI G, SQUILLARIO M, BARLA A, SANAVIA T, VERRI A, DI CAMILLO B (2012) **Discriminant functional gene groups identification with machine learning and prior knowledge.** In: 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN). April 25th-27th 2012, Bruges - Belgium.

[C6] SANAVIA T, CREPALDI A, BARLA A, DI CAMILLO B (2011) **Gene Ontology based classification improves prediction and gene signature interpretability.** Network Tools and Applications in Biology (NETTAB) Workshop, October 12th-14th 2011, Pavia, Italy.

[C7] SANAVIA T, AIOLLI F, DA SAN MARTINO G, BISOGNIN A, DI CAMILLO B (2011) **Improving biomarker list stability by integration of biological knowledge in the learning process.** In: 19th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 10th European Conference on Computational Biology (ECCB). July 15th-19 th 2011, Wien, Austria.

[C8] MINA M, SANAVIA T, DI CAMILLO B, TOFFOLO G, GUERRA C (2011) **Functional assessment of topological characterization using graphlet degrees in PPI networks.** In: 19th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 10th European Conference on Computational Biology (ECCB). July 15th-19 th 2011, Wien, Austria.

[C9] SANAVIA T, AIOLLI F, DA SAN MARTINO G, BISOGNIN A, DI CAMILLO B (2011) **Stable Feature Selection for Biomarker Discovery: Use of Biological Information.** In: BITS Annual Meeting 2011, June 20th-22nd 2011, Pisa, Italy. ORAL PRESENTATION

[C10] FACCHINETTI A, SANAVIA T, DI CAMILLO B, LAVEZZO E, FONTANA P, TOPPO S (2011) **A Method to Reveal and Handle Heterogeneities and Inconsistencies in Gene Ontology Annotation.** In: BITS Annual Meeting 2011, June 20th-22nd 2011, Pisa, Italy.

[C11] SANAVIA T, SAMBO F, GRASSI A, DI CAMILLO B, TOFFOLO G (2010) **Gene Network Inference by significance analysis on genotype/phenotype data.** In: 3rd Annual Joint Conference on Systems Biology, Regulatory Genomics and Reverse Engineering Challenges. Columbia University, New York, USA, November 16th-20th 2010.

[C12] SANAVIA T, FACCHINETTI A, DI CAMILLO B, TOFFOLO G, LAVEZZO E, TOPPO S, FONTANA P (2010) **Revealing heterogeneities and inconsistencies in protein functional annotations.** In: 9th European Conference on Computational Biology (ECCB). September 26th-29th 2010, Ghent, Belgium.

[C13] DI CAMILLO B, MARTINI M, SANAVIA T, JURMAN G, SAMBO F, BARLA A, SQUILLARIO M, FURLANELLO C, TOFFOLO G, COBELLI C (2010) **Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment.** In: 9th European Conference on Computational Biology (ECCB). September 26th-29th 2010, Ghent, Belgium.

[C15] SANAVIA T, BARLA A, DI CAMILLO B, MOSCI S, TOFFOLO G (2009) **Function-based analysis of microarray data via l1-l2 regularization.** In: 17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 8th European Conference on Computational Biology (ECCB). 27th June - 2nd July 2009 Stockholm, Sweden.

[C16] SANAVIA T, DI CAMILLO B, IORI E, MARAN A, BRONTE E, AVOGARO A, TOFFOLO G, COBELLI C (2008). **Function-based discovery of characteristic temporal expression profiles in endothelial cells stimulated with insulin.** In: 11th International Meeting of Microarray and Gene Expression Data Society (MGED). September 1st-4th 2008, Riva del Garda (TN), Italy.

National conferences

[C17] SANAVIA T, FINOTELLO F, DI CAMILLO B (2014) **FunPat: a function-based pattern analysis framework for RNA-seq time-series data.** In: IV Congresso Nazionale di Bioingegneria (GNB). July 25th-27th 2014, Pavia, Italy.

[C18] TREVISAN M, SANAVIA T, ALBONETTI C, LAVEZZO E, DI CAMILLO B, SINIGAGLIA A, TOPPO S, TOFFOLO G, COBELLI C, PALÙ G, BARZON L (2012) **Human cytomegalovirus microRNAs target prediction by dynamic expression analysis.** In: Terzo Congresso Gruppo Nazionale di Bioingegneria (GNB). June 26th-29th 2012, Rome, Italy.

[C19] DI CAMILLO B, MARTINI M, SANAVIA T, COBELLI C, TOFFOLO G (2010) **In silico assessment of effect of size and heterogeneity of samples on biomarker discovery.** In: Secondo Congresso Nazionale di Bioingegneria (GNB). July 8th-10th 2010, Turin, Italy.

Software

[S1] SANAVIA T, FINOTELLO F, DI CAMILLO B (2014) **FunPat: Function-based Pattern analysis on RNA-seq time series.** R package: http://sysbiobig.dei.unipd.it/?q=node/79

Bibliometric indicators

(from Scopus)

Documents: 8

Citations: 94

***h*-index:** 4

Honours and Awards

Grant for an oral presentation at 11th Annual Meeting of the Bioinformatics Italian Society (BITS). Rome, February 2014.

"Borsa Gini" Study Abroad Scholarship, awarded by "Fondazione Aldo Gini", Padova, Italy. Padova, January 2013.

Honourable Mention for Best Performer (fourth place) in DREAM5 (Dialogue for Reverse Engineering Assessments and Methods), Systems Genetics Challenge, Columbia University, New York, USA, November 2010.

National Award for Best Master Thesis in Bioengineering, Bressanone, Italy, September 2008.

Research activity

High-throughput technologies (e.g. Next-generation sequencing and microarray) are able to perform a high number of genetic tests simultaneously (e.g. monitoring of gene expression levels) which requires bioinformatics skills for the analysis. The research activity focuses on the development of computational methods able to improve the performance of the state-of-the-art methodologies for the analysis of both static and dynamic gene expression data, using prior knowledge derived from public genomic databases. Aim of the research is the robust identification of genes, biological functions and related regulatory mechanisms characterizing different phenotypes (i.e. observable physiological or pathological traits) or specific temporal genetic patterns, which is a fundamental step for early diagnosis and treatment. In particular, the following issues are/have been explored:

Integrative epigenomic and genomic computational methods for high-throughput sequencing data

Gene factors are insufficient in explaining all the aspects of heritable changes in biological functions, since there are mechanisms modulating the transcriptional regulation, such as epigenetic processes, which affect the genetic control of cellular processes. Sequencing techniques (e.g. RNA-seq, bisulfite sequencing and Chip-Seq) provide a huge amount of data, requiring computational methods for both identifying biomarkers characterizing a specific phenotype and integrating different sequencing data-types. However, these data are characterized by 1) strong differences between the number of variables and available samples, generating highly undetermined systems; 2) high correlations among variables (e.g. genes, epigenomic features). Recently developed machine-learning approaches which integrate prior knowledge from functional annotations of genomic databases have proven to be less sensible to data size and correlation issues by improving the stability in the identification of genomic markers. Current results are based on previous high-throughput techniques and performed on a single type of data, thus new efforts are required to both extend these approaches to sequencing data and

develop methods able to perform simultaneous analysis of different data types. Aim of the current research project is the development of new integrative approaches for the analysis of epigenomic and genomic sequencing data using knowledge-driven classification methods, in order to investigate if the combination of epigenome-wide regulatory markers with gene expression levels would provide greater power in detecting different traits and common patterns in biological systems, providing new insights into their complex functional relationships. Different approaches are currently being explored and evaluated in terms phenotype prediction, pattern stability and biological interpretability.

Function-based discovery of significant transcriptional temporal patterns
A method was developed to characterize temporal patterns from gene expression data. The algorithm exploits functional gene annotations from publicly available genomic databases and searches the main temporal patterns in classes of functionally related genes by using expectation maximization and weighted least squares. The output of the computational pipeline is a list of functional groups, each characterized by different temporal patterns and corresponding lists of differentially expressed genes. Tests on simulated data indicate a good performance of the method, in terms of both precision and recall. Application to real case studies has proven an improved biological interpretability of the results. Starting from 2012, applications of the method to new high-throughput sequencing data have been investigated focusing on the development and maturation of pancreatic beta cells, in collaboration with the Perelman School of Medicine at University of Pennsylvania (CBIL-Stoeckert Lab) and the Center for Stem Cell Biology at Vanderbilt University (Magnuson Lab). An R package for the analysis of RNA-seq time series expression data is now available [S1].
Scientific communications: 3 peer-reviewed articles reporting a preliminary analysis [J8], the description of the method and its application to microarray [J7] and RNA-seq [J1] data; 4 contributions to international [C1,C16] and national [C17-18] conferences.

Integration of prior information in supervised classification methods
Two integrative approaches were developed investigating the effect of different types of prior knowledge from genomic databases in supervised classification methods. In the first, biological databases on gene functions (e.g. Gene Ontology) and protein-protein interactions were considered to define functional groups of genes/proteins, defining a set of classifiers depending by these functional groups. The second approach codifies the biological information into similarity matrices which are integrated into kernel functions to be used in the learning process. The developed methods have proven an increasing stability in the identification of biomarkers and an improved biological interpretability of the results.
Scientific communications: 2 peer-reviewed articles in [J3,J5] presenting the two integrative approaches; 7 contributions to international conferences [C2,C5-9,C15].

Evaluation of supervised classification methods in high-throughput data
An assessment of how high dimensionality of high-throughput data and the typical within-class variability of measurements can affect classification performance and the identification of genes characterizing a specific phenotype (i.e. biomarkers) was performed. Bootstrap approaches applied on a range of state-of-the-art classifiers has been evaluated in terms of prediction performance, stability of biomarker lists and precision in biomarker selection, using both simulated and real datasets.
Scientific communications: 1 peer-reviewed paper [J6] describing the assessment on gene expression data and 1 peer-reviewed paper [J4] as part of an international consortium for a related work on critical assessment of classification methods in high-throughput data from flow-cytometry techniques; 2 contributions to international and national conferences [C13,C19].

Systems genetics approaches to understand complex genetic traits
Systems genetics refers to the study of complex genetic traits from high-throughput genetic and phenotypic data, using technologies such as gene expression arrays and sequencing. In this context, genetic variations can be thought as randomized, multifactorial perturbations of the gene expression profiles of each individual as the system response to a specific set of perturbations. A first integrative reverse-engineering approach was developed, which exploits both genetic and expression data to infer cause-effect relationships between genes. The obtained results provide

a basis for advanced integrative approaches able to automate the systematic interpretation of the gene regulatory network exploiting prior knowledge related to environmental factors and other important regulatory mechanisms such as epigenetic processes (e.g. methylation, histone modifications).
Scientific communications: 1 book chapter [B1] and a contribution to an international conference, which awarded an honorable mention from DREAM scientific community (Dialogue for Reverse Engineering Assessments and Methods) [C11].

Revealing and handling heterogeneities in Gene Ontology annotations
Large-scale sequencing projects have provided a huge amount of data related to gene and protein sequences. The more the number of genome/proteins sequenced, the greater the number of experiments performed to associate them functions. Most of the current functions are derived by computational prediction based on functional annotations from public databases such as Gene Ontology (GO). However, these predictions are known to be heterogeneous and sometimes inconsistent. Investigation of the presence of inconsistencies and heterogeneities in functional annotations was performed, developing a methodology which, by combining semantic similarity measures in GO annotations and sequence similarity between proteins, is able to identify outliers in clusters of proteins. Preliminary results of the application of this new methodology highlighted an elevated presence of heterogeneities with interesting distributions across GO annotations, suggesting useful information to handle its functional annotations.
Scientific communications: 4 contributions to international conferences [C3-4,C10, C12].

Seminars

Function-based discovery of temporal patterns in beta cell development. Genetics Research Talks. January 11th 2013, Perelman School of Medicine at University of Pennsylvania (Philadelphia, USA).

Attended conferences workshops and schools

Workshop **RNA-Seq Workshop for the Bioinformatician**, Università degli Studi di Milano, Dip. di Biotecnologie Mediche e Medicina Traslazionale, June 11th 2014, Milano (Italy).

11th Annual **Meeting of the Bioinformatics Italian Society** (BITS), Rome (Italy), February 26th-28th 2014.

Workshop **Computational tools for next-generation sequencing data analysis**, Fondazione Bruno Kessler, February 13th 2012, Povo, Trento (Italy).

11th International Workshop NETTAB: **Network Tools and Application in Biology**, Pavia (Italy), October 12th-14th 2011.

XXX Scuola di Bioingegneria, **Neuroinformatica**, Bressanone (Italy), September 19th-23th 2011.

PhD – **Summer School on Algorithms and Architectures for Computational Science and Engineering** (AACSE), Padova (Italy), September 12th-16th 2011.

19th Annual **International Conference on Intelligent Systems for Molecular Biology** (ISMB) and 10th **European Conference on Computational Biology** (ECCB), Wien (Austria), July 17th-19th 2011.

AFP/CAFA: The Automated Function Prediction SIG featuring CAFA: **Critical Assessment of Function Annotation**, Wien (Austria), July 15th-16th 2011.

8th Annual **Meeting of the Bioinformatics Italian Society** (BITS), Pisa (Italy), June 20th-22th 2011.

9th **European Conference on Computational Biology** (ECCB), Ghent (Belgium), September 26th-29th 2010.

XXIX Scuola di Bioingegneria, **Biologia Sintetica**, Bressanone (Italy), September 13th-17th 2010.

17th Annual **International Conference on Intelligent Systems for Molecular Biology** (ISMB) & 8th **European Conference on Computational Biology** (ECCB), Stockholm (Sweden), June 27th – July 2nd 2009.

11th **International Meeting of the Microarray Gene Expression Data** (MGED), Riva Del Garda (Italy), September 1st-4th 2008.

XXVI Scuola di Bioingegneria, **Genomica e Proteomica Computazionale**, Bressanone (Italy), September 24th-28th 2007.

Teaching activity

Teaching activity as professor

Course "**Bioingegneria per la genomica**", A.A. 2014/2015 (20 hours)

Teaching activity as assistant

Course "**Elaborazione dei segnali biologici**" (Prof. Gianna Toffolo), A.A. 2009/2010: preparation of laboratories and frontal lessons (54 hours)

Course "**Bioingegneria per la genomica**" (Ph.D. Barbara Di Camillo), A.A. 2009/2010: preparation of laboratories and frontal lessons (16 hours)

Advisory activity

Co-advisor in two master degree thesis:

Aler Crepaldi, "Integrazione di annotazione funzionale nella classificazione di dati di espressione genica", A.A. 2010/2011

Roberta Mazzucco, "Analisi di correlazione tra geni per la selezione robusta di biomarcatori", A.A. 2010/2011

Languages

Italian (Mother tongue)
English (Professional working proficiency)
French (Elementary proficiency)

Computer skills

Programming Languages

R (Expert)
MATLAB/Octave (Expert)
Python (Advanced)
Perl (Intermediate)
Java (Intermediate)
SQL (Intermediate)
Pascal (Intermediate)

Bioinformatics Tools

Main **Bioconductor/CRAN packages** for analysis of microarray and sequencing data, e.g. edgeR, DEseq, limma, affy, AgiMicroRna (Expert)
Quality control and preprocessing tools: FastX, FastQC (Advanced)
Alignment tools: BWA, Bowtie, RUM, RSEM (Advanced)
Bisulfite-seq tools: Bismark, methylKit (Advanced)
Chip-Seq tools: MACS, Homer (Intermediate)