ERCIM "Alain Bensoussan" Fellowship Scientific Report

Fellow:	Vincenzo Lagani
Visited Location :	Institute of Computer Science (ICS), Foundation for Research and Technology – Hellas (FORTH), Heraklion, Greece
Duration of Visit:	01 October 2009 – 30 September 2010

I - Scientific activity

The Fellow's research focused on the designing and validation of new Causal Analysis (CA) algorithms, and on their application on biological/bioinformatics data. CA techniques in Computer Science is a relatively recent research area, belonging to the fields of Machine Learning and Statistics: CA algorithms are able to infer causal relationships directly from a set of observations, and the inferred models can be manually inspected by human experts for acquiring useful insights into the process that generated the data. In the context of CA the Fellow faced two specific problems:

1. *Extending the Max – Min Parent and Children (MMPC), for variable selection in high dimensional, survival data*

The Fellow designed, implemented and validated the Survival MMPC (SMMPC) algorithm, i.e. a new version of the MMPC algorithm able to analyze right censored data. MMPC is a feature selection algorithm based on the theory of Causal Bayesian Networks and Markov – Blanket variable selection, that has recently shown to perform exceptionally well for data classification tasks. Survival analysis studies the timing and the occurrence of events of interest; one characteristic of survival data is that they are often right censored, i.e. the exact time of event's occurrence is not known for a subset of subjects under study. Due to censorship, standard feature selection algorithms can not directly deal with survival data. The Fellow performed a wide validation of the new algorithm on six different high dimensional gene expression dataset, and the results demonstrated that SMMPC is able to statistically significantly outperform several other feature selection algorithms for survival data, while returning in average the smallest set of variables. The results have led to a journal publications (see publications below).

2. *Designing constraint based algorithm for learning from mixtures of experimental and observational data*

Experimental studies (e.g., randomized controlled trials) are considered the gold standard to establish a causal relation. For example, different genes may be knocked-out in order to identify the genes that are causing a phenotype. Obviously, the gene-expression profiles of cells with different knock-outs follow different distributions, which is also different from wild-type (no knock-outs) profiles. Given that the distributions are different, the data of these studies cannot be simply pooled together. There exist a handful of methods able to learn from mixtures of observational and experimental data, but all of them present one or more limitations, the presence of latent variables can not be modelled, or interactions among variables are assumed to be linear. The Fellow worked on designing new algorithms based on the Maximal Ancestral Graph formalism (extensions of Bayesian Networks that include latent variables) able to overcome the aforementioned limitations. The first experimentations provided encouraging results and a publication is under preparation.

3. Study the effect of transcript features to the process of aging in wild-type and mutated mice.
During his stay the Fellow had also the opportunity of analyzing a large, non-public, biological dataset, provided by the Institute of Molecular Biology & Biotechnology (IMBB) of the Host Institution (FORTH). The dataset is composed by time course gene expression data measured on mice cavies, and the aim of the study is to understand the influence of certain characteristics of transcripts to the aging process. The results of this study are currently under preparation for publication.

II- Publication(s) during your fellowship

Lagani V., Tsamardinos I.
Structure-based variable selection for survival data
Bioinformatics (2010) 26 (15): 1887 – 1894

Abstract:
Motivation: Variable selection is a typical approach used for molecular-signature and biomarker discovery; however, its application to survival data is often complicated by censored samples. We propose a new algorithm for variable selection suitable for the analysis of high-dimensional, right-censored data called Survival Max–Min Parents and Children (SMMPC). The algorithm is conceptually simple, scalable, based on the theory of Bayesian networks (BNs) and the Markov blanket and extends the corresponding algorithm (MMPC) for classification tasks. The selected variables have a structural interpretation: if *T* is the survival time (in general the time-to-event), SMMPC returns the variables adjacent to *T* in the BN representing the data distribution. The selected variables also have a causal interpretation that we discuss.
Results: We conduct an extensive empirical analysis of prototypical and state-of-the-art variable selection algorithms for survival data that are applicable to high-dimensional biological data. SMMPC selects on average the smallest variable subsets (less than a dozen per dataset), while statistically significantly outperforming all of the methods in the study returning a manageable number of genes that could be inspected by a human expert.

III -Attended Seminars, Workshops, and Conferences

None.

IV – Research Exchange Programme (12 month scheme)

First Exchange institute:	Norges teknisk-naturvitenskapelige universitet (NTNU), Trondheim, Norway
Exchange dates:	May 3-12, 2010
Research contact:	Finn Drabløs

The NTNU research group leaded by Prof. Finn Drabløs has a strong experience in transcription factors binding sites prediction studies, and in general in transcriptome analysis. During his stay the Fellow worked in close contact with both Prof. Drabløs and the other components of the group, having several profitable talks and discussions. In particular, the Fellow acquired precious knowledge about several aspects of transcriptome analysis:

- the biological processes underlying the regulation of genes transcription;

- the technical procedures (e.g. Chromatin immunoprecipitation) employed for producing transcriptomics data;
- the analytical methods for analyzing transcriptomics data and predict transcription factors binding sites.

The Fellow had also the opportunity of giving a public talk, for illustrating the main principles of Causal Analysis and their possible applications in the field of transcriptome analysis. The mutual exchange of ideas and information during the public talk allowed the establishment of a collaboration between the research group of the Fellow and the Department of Oncology of the St. Olav University Hospital. This collaboration is still ongoing and it is aimed at studying the biological process underlying the rare Mesothelioma cancer by applying Causal Analysis methods.

Second Exchange institute: Valtion Teknillinen Tutkimuskeskus (VTT), Espoo, Finland
Exchange dates: May 14-21, 2010
Research contact: Matej Orešič

During his stay at the second Exchange Institute the Fellow worked with the components of the QBIX (Quantitative Biology and Bioinformatics) laboratory. The main topic discussed by the Fellow with the QBIX members was the integration of data produced by different "omics" techniques. Genomics, transcriptomics, proteomics and metabolomics techniques are nowadays able to provide huge amount of data, which are usually analyzed in isolation for separately modelling different biological mechanisms. However, a complete understanding of the biological processes requires a more holistic approach, able to integrate several types of omics data. The members of QBIX group illustrated different interesting approaches for integrating different omics data to the Fellow, including a visual methodology based on graph theory. On the other hand the Fellow gave a talk about the possible application of Causal Analysis method for integrating heterogeneous type of data, and he collected feedback on how to apply his theoretical research in the context of omics data integration.