Master Thesis proposal
Inference of T cell gene regulatory networks and master regulators of diseases using ODEs

Claudio Altafini
Division of Automatic Control, Dept. of Electrical Engineering,
Linköping University, SE-58183, Sweden.
email: claudio.altafini@liu.se

Mika Gustafsson
Center for Individulised Medicine,
Linköping University, SE-58183, Sweden.
email: mika.gustafsson@liu.se

December 15, 2014

This proposal is for a Master thesis in the field of **Systems Biology.**

Background and aims: The group of Mika Gustafsson has recently found that T cell diseases have significant enrichment of shared miss-expressed and highly interacting genes (Gustafsson et al. Genome Medicine 2014), which are highly enriched for disease associated genes. In cancer, the Califano group (e.g. Chen et al. 2014) has inferred gene regulatory networks from gene expression variations in cancer patients and used these to infer differential activity of master transcriptional regulators, which have been validated as cancer drivers. The idea of this project is to do the same for T cell associated complex diseases.

Methods and Materials: We will use the linear expression model (e.g. Zampieri et al. Bioinformatics 2010), but adding the LASSO constraint for regularisation and subset selection (see for example Gustafsson et al. ANYAS 2009 or Christley et al PLoS One 2009). Fast implementations of the LASSO exists for MatLab (the fastest is probably `://web.stanford.edu/~hastie/glmnet_matlab/`). The LASSO solves the parameter identification for each gene individually and can perform simultaneous subset selection and shrinkage, which makes it fast and possible to solve in parallel problems and gives a robust sparse solution. In this problem, for each target gene we will test which of the transcription factors (TFs) that are predicted to bind in an active promoter of the target gene do so in the data. Mika Gustafsson will provide the student with 8-10 normalised public expression data sets from T cell diseases, where each data set consists of about 20 -100 patients and controls. Moreover, we will provide sequence based predicted TF-target relationships for T cells.

Potential results The analysis will be performed for each dataset separately, each consisting of about 20.000 genes and 1600 TFs. For each target gene we will typically have 10 candidate TFs, for which we will do the inference. After having inferred the networks we should show that they make sense when compared with in vitro data, and possibly also suggest new experiments. Possible validations could be based on existing 1) Chip-seq of some TFs in T cells, 2) siRNA of some TFs followed by microarrays, 3) comparison with time-series data, 4) cross-validation between the data sets. Having validated the network we could use the eigenvectors and eigenvalues to back-track the expression in disease conditions and thereby suggesting the MR of each disease. We can then search for enrichment of DNA disease associations near the MRs, e.g. disease associated SNPs and differential methylation. Moreover if we really believe in the model we can use it to understand our existing dynamical data of disease and treatment.