

Abstract

Unravelling the molecular basis underlying genetic disorders is crucial in order to develop effective treatments to tackle these diseases. For many years, scientists have explored which genetic factors were associated with several human traits and diseases. After the completion of the human genome project, several high-throughput technologies have been designed and widely used, therefore producing large amounts of genomic data. At the same time, computational tools have been developed and used in conjunction with wet-lab tools to analyze this data in order to enrich our knowledge of genetics and biology.

The main focus of this thesis is gene prioritization, that can be defined as the identification of the most promising genes among a list of candidate genes with respect to a biological process of interest. It is a problem for which large quantities of data have to be manipulated, which typically means that it has to be done *in silico*. This thesis describes two gene prioritization methods from their theoretical development to their applications to real biological questions.

The first part of this thesis describes the development of two data fusion algorithms for gene prioritization respectively based on order statistics and kernel methods. These algorithms have been developed for human and also for reference organisms. Ultimately, a cross-species version of these algorithms have been developed and implemented. Integrating genomic data among closely related organisms is relevant since many researchers are studying human indirectly through the study of reference organisms such as mouse or rat, and are therefore producing mouse/rat specific data, that is still relevant in human biology. Our method can integrate more than 20 distinct genomic data sources for five organisms and is therefore one of the first cross-species gene prioritization method of that scale.

Only a fragment of all the computational tools developed each year specifically for biology are still maintained after three years, and even less are used by independent researchers. The second part of this thesis focuses on the benchmarks of the proposed methods, the development of the corresponding web based softwares, and on their application to real biological questions. By making our methods publicly available, we make sure that interested users can apply them for their

own problems. In addition, benchmarking is needed to prove that the approach is theoretically valid and can estimate how accurate are the predictions. Ultimately, the inclusion of our computational method within wet-lab workflows show the real usefulness of the approach.