Hong Kong University of Science and Technology
Division of Biomedical Engineering

PhD Thesis Presentation
Bioengineering Graduate Program (BIEN)

Qualitative and Quantitative Protein Interaction Prediction with Machine Learning

by
MARINI Simone

Abstract
Protein interactions constitute a crucial parts of the cell machinery. In particular, protein-protein, DNA-protein and RNA-protein Interactions have a critical role in the whole cellular system of transcription-expression-regulation. Machine Learning is utilized to predict protein interactions through empirical models: simulations are viable tools to avoid time-spending, resource-consuming experiments, which also require expensive equipment and trained biologists to be performed. Furthermore, Machine Learning-based models are not influenced by human bias and may suggest unexpected research paths. There are two kinds of predictions: (a) qualitative interaction prediction about the interaction existence (classification), and (b) quantitative affinity prediction about the interaction affinity. By predicting the presence or the absence of the interaction, we perform a qualitative prediction. Although many progresses have been achieved in recent developments, the accuracy and reliability of qualitative protein-protein, DNA-protein and RNA-protein interaction classifiers are far from the state that could be potentially achieved. In this work, we provide viable tools for protein interaction and protein affinity prediction. We present novel approaches for protein-protein interaction prediction, DNA- and RNA-protein interaction prediction (involving aptamers) and quantitative protein-protein affinity prediction in the molecular context of Dscam proteins. The proposed protein-protein interaction classifier is based on a ensemble of classifiers. By combining different techniques and descriptors, this ensemble classifier overcomes the flaws of single algorithms composing it. Compared to previously published works trained on the very same data it provides more accurate results. Our approach for the DNA- and RNA-protein interaction classification considers the problem from the perspective of protein pairs. This perspective allows to embed multiple targets in our data sets providing pair-specific predictions. We present four different ways to assemble negative (non interacting) instances, and we measure how each method affects the results. We then use it predict new putative interacting aptamers. The quantitative affinity prediction method is tailored on Dscam protein-protein Affinity prediction. Despite Dscam is an important protein family, critical for neural development, Dscam affinity prediction has never been attempted before. The model provides predictions about thousands of self-binding Dscam proteins, while its feature ranking allows to investigate the evolution of Dscam proteins binding machinery.

Date : **3 Aug 2012 (Friday)**
Time : **11:00am**
Venue : **Room 4504 (Lift 25-26)**
All are welcome!

Examination Committee:
Prof. Jianhua Chen (Chair)
Prof. Qiang Yang and Hong Xue (Supervisors)
Prof. Henry Lam
Prof. Ying Chau
Prof. Francesco Ciucci
Prof. Lei Zhang (HK Poly U)
Prof. Michael Ng (HK Baptist U)