若手研究者インターナショナル・トレーニング・プログラム(ITP)

バイオインフォマティクスとシステムズバイオロジーの国際連携教育研究プログラム 応募書類

Name: Sayaka Mizutani

Title: Statistical analysis of breast cancer progression based on genomic, transcriptomic and phenotypic data

Institute: Bioinformatics Center, Institute for Chemical Research, Kyoto University

Partner institute of your choice：Centre for Computational Biology, Mines ParisTech

Duration of your choice: September 30, 2010 – December 24, 2010

Research Objectives

I will be dedicated to a cancer research under the supervision of Dr. Jean-Philippe Vert and Dr. Yoshihiro Yamanishi at Centre for Computational Biology, Mines ParisTech.

Centre for Computational Biology is dedicated to epidemiology, bioinformatics and systems biology of cancer. Their research objective is the development of probabilistic models and statistical machine learning to process and analyze data produced by high-throughput technologies. The laboratory has a joint program with Institut Curie**,** one of the leading medical, biological and biophysical research centers in the world. The two institutes have established cooperative research projects, in which the developed computational models are applied to experimental data obtained from experimental laboratories in Institut Curie. Staying in Mines ParisTech would be a valuable experience in working both with researchers in computational biology and experimental biology. Therefore, I would like to apply for the ITP program for three-month stay in Mines ParisTech.

There has been an increasing demand for bioinformatics approaches in the analyses of high-throughput data, which vary from genetic polymorphisms, to gene copy numbers, to microarray gene expression, and to data produced by next-generation sequencers. Accordingly, several research laboratories including the Centre of Computational Biology have been dedicated to developing computational methods to extract useful information from these data sets. One of the goals of my collaborative research is to learn computational methods, which have been studied by the laboratory and been applied to the analyses of several medical and biological data. Another important role of bioinformatics in medical research is to collect biologically relevant knowledge in relation of clinical symptoms. Kanehisa laboratory has been developing the KEGG database, a widely accepted database in which molecular biological information regarding to several cancer pathogenesis are accumulated. My participation to the three-month program will provide an opportunity to further develop a collaborative partnership between two laboratories by integrating the resources of the two laboratories. Finally, I will anticipate in interacting with researchers in other clinical laboratories in the Institut Curie. The Centre for Computational Biology participates a joint seminar with many laboratories in the institute. By participating discussions with researchers from different clinical interests will provide me many opportunities to learn the essence of clinical research in a new viewpoint.

Research Plan

1. Background

Breast cancer is one of the most common types of non-skin cancer in women and the fifth most common cause of cancer death. Recent advances in the analyses of high-throughput data have enabled us to classify breast cancers into different subtypes. However, the mechanism of cancer progression has not been well understood. Unlike the colon carcinomas, which are well explained by the Volgenstein's adenoma to more malignant carcinoma progression model, the complexities in the pathological entities of breast cancer makes it more challenging to set up a good model to understand cancer progression. Recent genotypic-phenotypic correlation analyses have delineated evolutional pathways, which lead to different subtypes of clinical behaviors. However, our knowledge on molecular alterations associated with each step of progression is only fragmental. In my research we will analyze recently published breast cancer data to identify stage-specific alterations of genes by applying a statistical method called the canonical correlation analysis (CCA).

2. Materials and methods

I will analyze a breast cancer data published by Chin *et al.* in 2006 [1]. It consists of three types of data; CGH, gene expression, and clinical diagnostic data taken from the same set of patients. CGH covers gene copy number alterations, which have recently shown to be causative to the malignancy of breast cancer. Gene expression data covers the aberrations of mRNAs. Clinical diagnostic data includes the histopathological grades of the tumors, the results of cancer markers and other features of the patients.

I will apply a recently developed version of the canonical correlation analysis (CCA) to the Chin *et al.* data. Given two sets of data over the same sets of samples, CCA finds elements that maximize the correlation between the two data sets. A useful algorithm has been provided by Witten *et al.*, 2009 in an implementation to the widely used software R [2]. Using this algorithm we look for sets of genes that are altered both in the gene copy numbers and the transcription levels. Then, we classify gene sets into different subtypes of breast cancer in the progressive stages and clinical outcomes. This approach will help us predict genomic and transcriptomic alterations of genes that may give rise to a specific stage or a specific clinical of the tumor.

3. References

[1] Chin, K. et al., Genomic and transcriptional aberrations linked to breast cancer pathophysiologies, *Cancer Cell*, 10:529–541, 2006.

[2] Witten, D.M., Tibshirani, R., and Hastie T., A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics*, 10:515–534, 2009.