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 報告書

Name：Sayaka Mizutani

Title：Analysis of the correlation between drug side-effects and drug-target interactions.

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

Partner institute: Centre for Computational Biology, Mines ParisTech

Duration: September 30, 2010 – December 24, 2010

Report:

1. General report

Supported by the ITP program, I stayed for three month at Mines ParisTech, Centre for Computational Biology, directed by Dr. Jean-Philippe Vert. During the stay I was involved in a research project under the supervision of Dr. Yoshihiro Yamanishi, a researcher at the laboratory. The Center for Computational Biology has a joint research program with Institut Curie. Institut Curie is one of the leading medical and biological research institutes in Paris and a clinical institute specialized in the treatment of cancer. Faculty members and many students in the Centre for Computational Biology run collaborative research projects with experimental biologists in the institute. In the projects they develop theoretical and computational framework for the analyses of large-scale data. Staying in Institut Curie would be a good choice for me to experience the atmosphere of collaborative research between computational biology and experimental biology. Therefore, I chose to stay at Institut Curie in Paris.

During the stay I participated in weekly lab meetings, where students and faculty members are responsible for a brief progress report. Every meeting was also devoted for a longer presentation time, in which one of the members is expected to give a detailed talk on their research projects. From the meetings I was exposed to many new concepts and ideas in the Machine Learning field. Also the meetings were perfect opportunities for me to exchange opinions with the lab members over each other's research topics. I also participated in a reading group on Statistics and Machine Learning topics.

In Institut Curie, there were open access seminars, which are organized by several different laboratories. These seminars cover many research fields including Molecular, Developmental Biologies, Oncology and other medical studies. Some of such seminars, that were impressive to me, were given by Dr. Edith Heard from Institut Curie, on the inactivation of X chromosome, and by Dr. Reinhard Laubenbacher from Virginia Bioinformatics Institute, on algebraic modeling of iron metabolism. There was also a talk given by Dr. Elizabeth Blackburn from University of California, San Francisco, the 2009 Nobel Prize laureate for the discovery of telomerase. These seminars were valuable opportunities to learn about interests and ideas from experimental

biologists, from which I could search for possibilities for computational biology methods to apply to problems that may occur in the analyses of large-scale data.

Furthermore, I had a few chances to listen to Statistical Machine Learning lectures at l'Ecole Normale Supérieure located a few minutes from Institut Curie. Especially it was a very valuable experience to listen to a lecture from Dr. Trevor Hastie, who is one of the most famous researchers in the Statistical Machine Learning field.

2. Research project

In the first two weeks of stay I worked on a breast cancer data published by Chin *et al.* in 2006 [1]. The dataset consists of three types of data; microarray expression, array CGH, and diagnostic data. Our goal was to extract genes whose expressions are correlated with the copy numbers of the corresponding chromosomal areas. In order to predict cancer progressive stages based on the extracted genes we performed a canonical correlation analysis (CCA) and a support vector machine. However, we found that the calculation would take several months to finishi, due to the large size of the data. Therefore, I had to change my research topic.

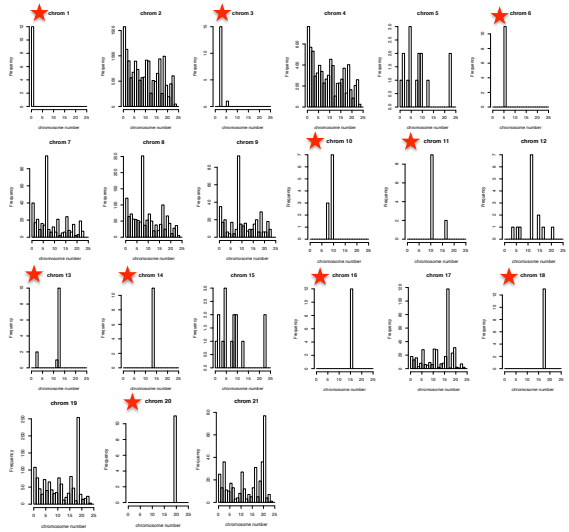


Figure 1. Histograms of the frequency of genes whose expressions are correlated with copy numbers. Red stars indicate chromosomes for which mostly true positives were gained.

Next, I became involved in a project organized by Dr. Yoshihiro Yamanishi and his colleagues, Dr. Véronique Stoven and Edouard Pauwels from Mines ParisTech. Their interest is to develop statistical models that can be applied to a system-wide analysis of multiple drug-related data of different structures. Drug-related database such as DrugBank [2] accumulates different types of drug-attributed data such as chemical structures, target proteins, and drug indications. Linking such different data may provide molecular and physiological explanations how drugs act in the human body. From such a point of view, Yamanishi and colleagues have been worked on the development of statistical machine learning tools to investigate relationships between different types of drug-attributed data. In previous work, they proposed a method to predict pharmacological effect information of drugs using their chemical structures [3]. However, this approach does not shed light on the mechanism in which drug-target protein interactions cause side-effects. Thus, we were motivated to investigate correlations between drug target proteins and their side-effects. Extraction of correlated sets of target proteins and side-effects is made possible by considering a CCA. Given two sets of variables from the same set of samples, CCA solves for the variables in which the correlation between the two sets of variables is maximized. Moreover, more recent version of the method has been proposed, in which only a few variables are extracted [4]. We applied these

methods to our datasets. Cross validation indicated that our methods have high prediction accuracies. We also validated the contents of extracted sets of target proteins and side-effects from a biological point of view. As a result, we came to a conclusion that our method can be applied to the prediction of potential side-effects for drugs based on their target protein profiles. Then, using target protein profiles of drugs whose side-effects are unknown, we predicted potential side-effects for these drugs. The result has been submitted to ISMB2011.

References:

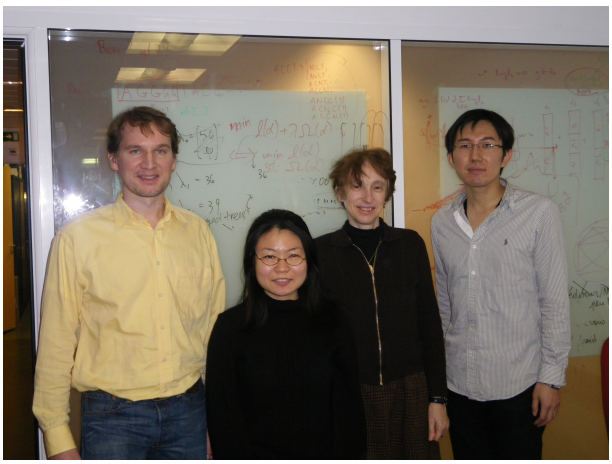
[1] Chin, K. et al., (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies, *Cancer Cell*, 10:529–541.

[2] Wishart, D., Knox, C., Guo, A., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. & Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, **34**, D668–D672.

[3] Yamanishi, Y., Kotera, M., Kanehisa, M. & Goto, S. (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, **26**, i246–i254.

[4] Witten, D., Tibshirani, R. & Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515-534.

Acknowledgement:
I would like to express my deepest gratitude to Professor Minoru Kanehisa and Professor Hiroshi Mamitsuka, and Associate Professor Susumu Goto for providing me a valuable learning experience through the ITP program. I would also thank Dr. Jean-Philippe Vert, Dr. Yoshihiro Yamanishi and their colleagues for kindest supports for my stay in Institut Curie.



From the left; Dr. Jean-Philippe Vert, Sayaka Mizutani, Dr. Véronique Stoven and Dr. Yoshihiro Yamanishi.

A snap shot from a weekly lab meeting at The Centre for Computational Biology.