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：NGUYEN HAO CANH
Title：Latent Feature Models for Biological Knowledge Discovery
Institute: 京都大学化学研究所バイオインフォマティクスセンター
Partner institute: Centre for Computational Biology, Mines ParisTech – Curie Institute – INSERM U900 joint laboratory, 11-13 rue Pierre et Marie Curie, 75005 Paris
Duration: Dec 15th 2012 ~ Mar 14th 2013
Report: Under the International Training Program (ITP), I had a chance to visit MINES ParisTech for three months. My destination was the Center for Computational Biology of the university, headed by Dr. Jean-Philipe Vert. The center is in a joint laboratory with Curie Institute (one of the leading medical, biological and biophysical research centers in the world) and INSERM (the French National Institute of Health and Medical Research). The laboratory is located in the vibrant 5th arrondissement of Paris in the Latin quarter, known for its student life, lively atmosphere and bistros. It is within a walking distance to the likes of the Luxembourg garden, the Pantheon, the Notre Dame Cathedral. It is a very interesting experience to stay and observe how it works in a laboratory of that inter-disciplinary nature. The laboratory gathers researchers from many different backgrounds that offer a very exciting chance to see researches ranging from Applied Mathematics, Statistics, Computer Science to Biology, Chemistry and Medical/Clinical Sciences. With the well-known hospital for treating cancer inside the campus, many works in the laboratory are exploratory data analysis for data coming directly from the hospital, in collaboration with medical doctors. On the other hand, inside the laboratory and with collaborators, there are many computer scientists, mathematicians around to work on the theory of statistical analysis, machine learning and optimization. While regular biweekly laboratory seminars are mainly about data analysis on real clinical data from the hospital, regular talks of visitors and weekly group talks are usually more on the theoretical sides. One of the things I enjoy the working environment is the openness of people to idea exchanges. It is of particular importance for inter-disciplinary researches. It is common to see that many projects involve many researchers from difference sides. Regular face-to-face discussions and skype meetings are the way to make sure that everyone is on the same page (lesson learnt: keep the cafeteria's door open and the coffee flowing). Within the laboratory, it is common to see people meet up to work out the maths or to double-check the computer programs together. Common activities are organized to serve two purposes: (i) to equip everybody with necessary background knowledge, specially on the theoretical sides, and (ii) to discover and build teams of similar interests. Regular weekly group reading on optimization provides new comers with background knowledge, guided by the more experienced. It is not just about reading and understanding the text. The same amount of time is spent on doing exercises and when necessary, coding and running the methods to actually experience the abstract concepts from theory to practice. I found this very useful and could be helpful should I embark on this direction. These are the particular experiences that I find useful, thanks to the ITP program.

Report (Continued)：

During my stay, I took the opportunity to work with mathematically-minded researchers to shift my research more into the side of principles of statistical methods. I worked on a project I initiated here, and tried to move forward through the interactions with the mathematicians there.

The problem that I am considering is of fundamental importance in modern data analysis with structured data, specifically graph data. Graphs are models of networks in many statistical analysis tasks. In Systems Biology, networks can be seen anywhere: metabolic pathways, interaction networks, regulatory networks, signaling networks and so on. Nowadays, high-throughput experiments and collective databases produce large datasets that require expensive computational models to make sense out of them. Statistical analysis usually becomes more reliable with more data available. However, for networks, the same thing can not be said. We worked a phenomenon reported recently that the larger the graphs, the less information can be extracted for many available statistical analysis methods.

Many methods for statistical analysis of graph data rely on the information in graph Laplacian to utilize graph structures. One common example is semi-supervised learning arisen in the situation where biological networks can be collected while labeling the nodes on the network, i.e. determining their functions, is too costly. Another example is the case in which biological mechanisms go through networks, therefore, graphs regularization for optimization methods are usually used. In these methods, one usually uses the graph Laplacians as the regularizers in their methods (such as Laplacian Support Vector Machines). Equivalently, the inverses of graph Laplacians are usually used as scores for similarity/distance (such as hitting/commute time distance) to feed into machine learning methods.

It was recently discovered that as the graphs become larger, the inverses of graph Laplacians tend to lose global information of graph structures, only local information is kept. The root of the problem is that, the information transferred in the graphs following Markov models does not go far. As the network becomes larger, the information becomes more localized, as opposed to the global graph structures are we were aiming for. This finding virtually renders many methods and applications with graph Laplacians not having a sound basis.

We studied the phenomenon and found that, walks on graphs corresponding to Markov chains contain both global information and local information of graph structures at the same time. Both types of information are averaged to make the final scores of similarity or distance for statistical methods. However, due to the fact that in large graphs, global information becomes an order of magnitude smaller. At certain scale of network, the difference in global information is absorbed in the difference in local information, making the graph Laplacians unusable. To account for this problem, we aim to design different random walks on graphs that emphasize more on the global information, i.e. having more weight on the non-local parts that they go through. We are working toward a general framework of weighted random walks to unify many different approaches on the same problem. We are also using the same analysis to shed light on various random walk models previously used.

Report (Continued)

A trip to Paris would not be complete without various extra-curricular activities such as hanging out with host fellows, enjoying museums and exhibitions, satisfying the curiosity with the long history of the country and the eat-everything culinary style. It served me very well with my photographic interest in the city of light, the surrounding beautiful countryside and neighboring European cities. Following is the report in picture.