Name：Peiying Ruan

Title：Prediction of heterodimeric protein complexes using protein-protein interaction networks

Institute: Bioinformatics Center (Akutsu Laboratory), Kyoto University Institute for Chemical Research; Kyoto, Uji, Gokasho 611-0011, Japan

Partner institute: Centre for Computational Biology, Mines ParisTech

Duration: September 4, 2013 – November 24, 2013

Report:

1. Introduction of Laboratory:

During the International Training Program, I was staying in Dr. Jean-Philippe's laboratory in Centre for Computational Biology building at Curie Institute (Figure 1: top), which belongs to Mines ParisTech. The members of several groups in the building share the rooms and computers. The computers and seats are free to be chosen since every member has a personal account to login all the computers. There are three postdoctoral fellows and about 10 PhD students in our group, including two new coming postdoctoral fellows at the time I was leaving. We have a basic weekly group meeting on every Tuesday and machine learning reading club on every Thursday. Besides that, we sometimes have smile seminars in other buildings to listen to others' talks (Figure 1: down). In the weekly group meeting, we first discuss some news on the lab things and then talk about current research and new progress one by one, and sometimes we can give presentations if we want to. Everyone is active and they always make jokes during talking, so we can totally enjoy the seminar in a relaxed mood.

Figure 1. Environment of the lab

Curie institute locates in the central position in Paris, with Seine River and Luxembourg Garden around. Therefore, I could enjoy a nice view everyday when I went to the laboratory. I was living outside Paris during my stay. It took me near 50 minutes to go to the lab (train+walk). But I really enjoy the time when I was taking the train. There are nice views outside and sometimes people play music in the train. While listening to the music, I looked at the beautiful scenery outside and felt peaceful. It was a good time to think about the research, even sometimes a new idea could come out.

2. Research project:

The research interest of Dr. Jean-Philippe mainly concerns the development of statistical and machine learning methods for computational biology. My previous research used kernel

Report (Continued)：

functions, and Dr. Jean-Philippe is an expert in this area. So we considered continuing my previous research to improve the results. My previous work is on prediction of heterodimeric protein complexes using protein-protein interaction networks. As we know, protein complexes play crucial roles in a variety of biological processes, such as ribosomes for protein biosynthesis, molecular transmission and evolution of interactions between proteins. In fact, many proteins come to be functional only after they interact with their specific partners and are assembled into protein complexes. Hence, much effort has been made for predicting protein complexes from protein-protein interaction (PPI) networks in bioinformatics, such as MCL, MCODE, RNSC, PCP, RRW, and NWE.

These methods have dealt with only complexes with size of more than three because the methods often are based on some density of subgraphs. However, heterodimeric protein complexes that consist of two distinct proteins occupy a large part according to several comprehensive databases of known complexes. In our previous work, we proposed several feature space mappings from protein-protein interaction data, in which each interaction is weighted based on reliability. Furthermore, we made use of prior knowledge on protein domains to develop feature space mappings, domain composition kernel and its combination kernel with our proposed features.

The results suggest that our proposed kernel considerably outperforms other methods for predicting heterodimeric protein complexes.

Through the discuss with Dr. Jean-Philippe, we improved the performance for prediction heterodimers by the following steps:

(1) Using Kernels: Domain composition kernel that we proposed in the previous work is a binary kernel related to domain composition. We wonder if some existing pairwise kernels may perform better by comparing domain composition of each two complexes, such as tensor product pairwise kernel (TPPK) (1), metric learning pairwise kernel (MLPK) (2) and Tanimoto Kernel (3). **The results indicate that MLPK improved the prediction accuracy, while TPPK and Tanimoto Kernel did not.**

$$K_{KTTP}\left((x_1,x_2),\ (x_3,x_4)\right) = K_g(x_1,x_3)K_g(x_2,x_4) + K_g(x_1,x_4)K_g(x_2,x_3) \quad (1)$$

$$K_{MLPK}\left((x_1,x_2),\ (x_3,x_4)\right) = \left(K_g(x_1,x_3) - K_g(x_1,x_4) - K_g(x_2,x_3) + K_g(x_2,x_4)\right)^2 \quad (2)$$

$$K^{Tanimoto}(x_1,x_2) = \frac{K(x_1,x_2)}{K(x_1,x_1) + K(x_2,x_2) - K(x_1,x_2)} \quad (3)$$

(2) Add other information: In the previous work, the features we designed mostly based on the interactions between protein pairs. However, much other information such as sequence, expression, localization and phylogenetic profile, is not considered yet. Therefore, we collected these data from databases and employed proper kernel function for each (Figure 2).

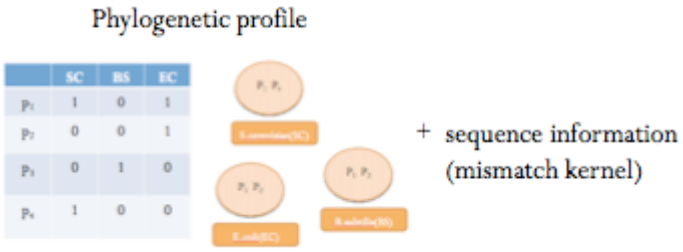

Figure 2: add other information

(3) Combine multiple kernels: Construct a kernel model to combine these kernels, with optimal weighting coefficients for each base kernel.

(4) Design a new kernel: We have proposed domain composition kernel, which improved the prediction accuracy a lot. However, domain composition kernel is a simple binary kernel. Hence, if possible, we will try to design a new kernel with more information.

(5) Experimental results & Discussion:

We performed ten-fold cross-validation computational experiments. By using the information and kernels mentioned above, and adding the feature mappings we made in the previous work, we obtained the results in the following table. Note that the best accuracy, evaluated by F-measure, we achieved in **the previous work is 0.631**. The first row is the result obtained by the method [Feature mappings + TPPK kernel], we can see that the result was not improved. But by the method [Feature mappings + MLPK kernel], the accuracy was improved to 0.656. In addition, we (i) added the expression information, the result became worse,
(ii) added the phylogenetic profile, the result was improved.
Then, we added sequence information and phylogenetic profile together, **the result was further improved up to 0.672** (Table 1).

Method	F-measure
Feature mappings +TPPK	0.605
Feature mappings + MLPK	0.656
Feature mappings + MLPK +Expression data	0.526
Feature mappings + MLPK + Phylogenetic profile	0.663
Feature mappings + MLPK + Phylogenetic profile+Sequence	0.672

Table 1: results of combined methods

3. Other activities:

Besides the research, we also had some parties during my stay in Paris (Figure 3: down left and right), I really enjoy the food there, especially like various French cakes. All the members

Plan (Continued)

in the group are very kind and offered me a lot of help. It's really hard to say goodbye with them. On my last weekday in Paris, we took a picture together (Figure 4). We wanted to take picture outside the building, but Dr. Jean-Philippe suggested us to take a picture in the laboratory, he smiled and said, "because we are always working", making us laugh out loudly. In my spare time, I visited Louvre Museum (Figure 3: top left), Eiffel Tower (Figure 3: top right) and many other famous architectures.

Figure 3. Some activities

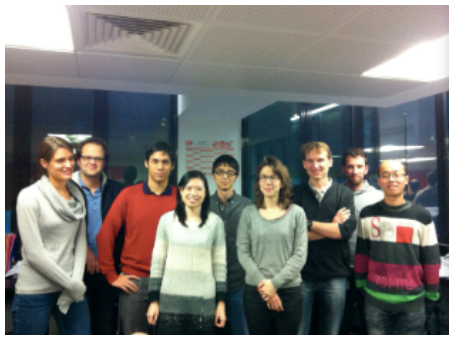

Figure 4. Members in the lab

Acknowledgement

First of all, I would like to express my thanks to my supervisor Professor Akutsu for allowing me participating this program, program director Professor Mamitsuka and Professor Kanehisa for providing me such a valuable opportunity. Also, I would like to thank Professor Jean-Philippe for giving me so many valuable advices about my research and all the group members for their kindness and supports during my stay in Paris.