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: Peiying Ruan

Title: Prediction of protein complex structures using neural networks

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

Partner institute: Macromolecular Modelling Group (Knapp Laboratory), Free University of Berlin; Fabeckstrassen 36a, D-14195, Germany, Berlin-Dahlem

Duration : April 8, 2011~July 5, 2011

Report :

Introduction of laboratory

During the international training program, I was staying in Knapp laboratory at Free University of Berlin, whose research is mainly about protein structure prediction, calculation of free energy of protein association and so on. Knapp laboratory is a laboratory in a separated small building from other laboratories. Professor Knapp has his own office while two or more other members share one room. I was sharing a room with two Ph.D., one of whom is my collaborator. The computer environment in Knapp lab is good enough for them to run programs they need. They run their own Linux cluster which is continuously upgraded. Currently, they have *tens of* dual and single CPU PC's (x86 and AMD64/Opteron) with up to 3.0 GHz connected simultaneously with Gbit switches. They hold seminar once a week, also attend other seminars from time to time. All of the Knapp lab members are very nice. They helped me a lot when I first reached this strange place that I felt so warm. It takes about 30 minutes to go to the lab (bus + train + walk) from the place where I was living.

Figure1. The road I walked everyday in FU. The house I was living in.

Research objectives

In Knapp laboratory, the research I was working on is based on a defined project, which is about protein-protein docking. What I was supposed to do was to improve that program. First, I would like to introduce some related work. Figure 1 shows the overview of the knowledge-based potentials. There are 191 Training protein-protein complexes (48 complexes from Benchmark 3.0[1] and 143 complexes from Huang et. al[2]) and 2000 near-native decoys

atom type	*description*
C2S	side chain carbonyl carbon
Nar	aromatic nitrogen
S31	sulfur with one hydrogen
Car	aromatic carbon
...	
HNi	hydrogen bond to nitrogen
HOS	hydrogen bond to oxygen or sulfur

Figure2. Atoms contacts

(Report Continued):

($0 \sim 6 \mathring{A}$ interface RMSD) for each complex of training set. For atoms contacts (Figure 2), 22 different atom types (20 heavy atoms and 2 hydrogen atoms) (Figure 3) were used to contact. Also, contacts were separated into 8 distance bins. Therefore, there were exactly $\frac{22*23}{2}*8=2024$ classes of contacts in total. Each decoy was described by its number of contacts for each class using vector of size 2024. As the last subfigure of Figure4 showed, one neuron standed for each of the 2024 contact classes. Then, train neural network using back-propagation to minimize error between calculated and desired output.

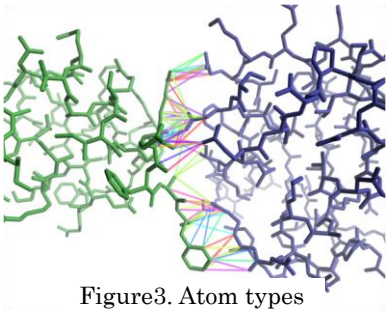

Figure3. Atom types

65 protein complexes from Protein Docking Benchmark 3.0[1] set were used as prediction data. Also, ZDOCK 3.0[3] provided 54,000 decoys for each complex. Decoys used the unbound structures of the proteins. For each decoy, they computed contacts and used trained neural network to calculate score.

Figure 4 shows the overview of the knowledge-based potentials, which is one of the scoring functions.

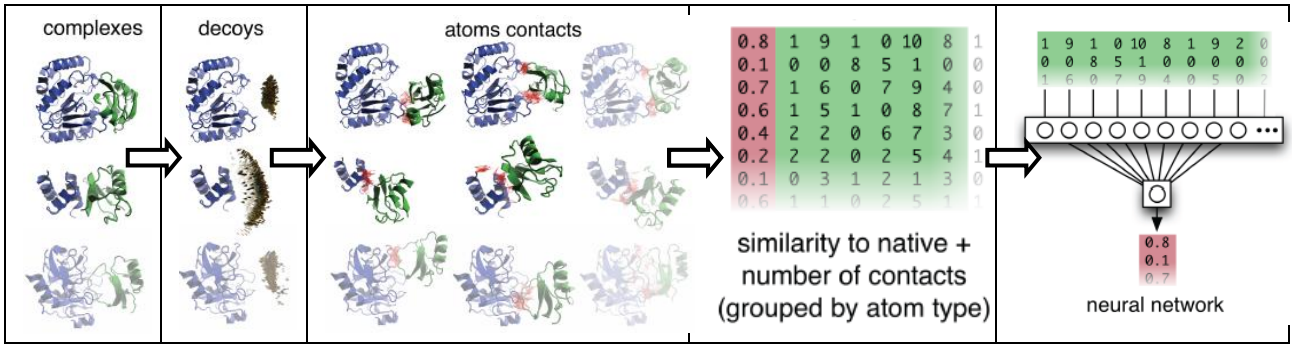

Figure4. Overview of the knowledge-based potentials

Our goal is to improve the program to increase the success rate for prediction.The detail work of what I was working on can be divided into 4 points below.

(1) Find good distance classes.

I tried 20 groups of different distance classes, each of them took almost one day to run. And finally I found several good distance classes, which have increased the successful cases of prediction. Figure 5 shows one of the best distance classes. The number of successful cases is 12 when the number of predictions is 1. Our goal is to reach 14.

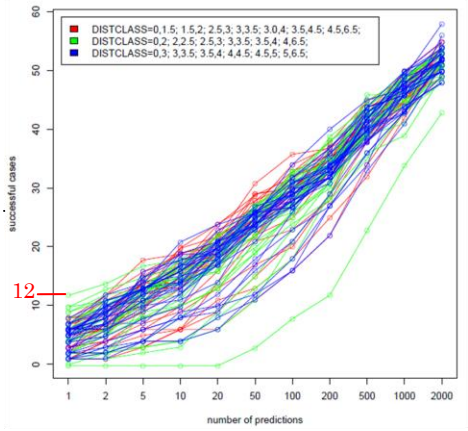

Figure5. One of the best distance classes.

(2) Analyze atom classes for the amino acids.

I translated atoms within PDB files of bound complexes into corresponding atom type which is show in Figure2.

(3) Analyze distance classes:

- Extend the range of distance class to $9 \mathring{A}$. $\Rightarrow$ Increase successful cases
- Combine distance classes for hydrogens.

$$\frac{22*23}{2}*8 \Rightarrow 1+\left(\frac{22*22}{2}+\frac{22}{2}\right)*(8-1) \Rightarrow \text{Take less time to calculate}$$

(4) New method to compare performance of two predictions.

It's hard to tell which method is the best one when looking at the results in the existing paper. To deal with the problem, I tried both Cumulative Hypergeometric Distribution (CHD) and cumulative distribution (CD). CHD took much more time to run than CD, at the same time, CD performances better than CHD in many cases. So, I used CD

$$prob = 1-\left(1-\frac{NN2.5}{total}\right)^{rank}$$

to calculate the probabilities of finding at least one of the [*NN2.5*] with [*rank*] tries out of [*total*] for each method. Then average each of the probabilities, as the final values to compare the methods. The lower the value is, the better the method is.

Then I applied this new method to compare the program I have improved (neural network-Peiying) with the previous program (neural network-Chae) and another pretty good program (ZRANK). The result shows that neural network-Peiying is better than other two methods (the lower, the better).

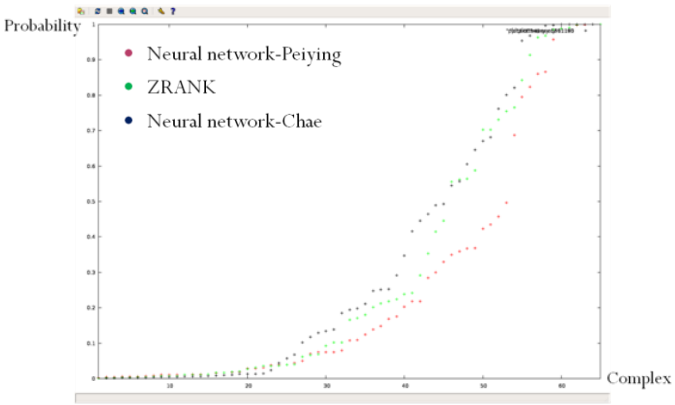

Figure6. Compare three methods (sorted)

Other Activities

During my stay in Free University of Berlin, I joined them in eating super big pizza (top-left), experienced the graduation party of Ph.D. (top-right), visited the Berlin Wall (down-left), attended the Carnival of Cultures in Berlin (down-right) and also travelled to other cities.




Acknowledgements

Finally, I want to express my thanks to my supervisor Prof. Akutsu for allowing me participating this program, program director Prof. Mamitsuka and Prof.Kanehisa for giving me this chance, also thanks to associate Prof. Ernst-Walter Knapp for giving me so many advices about my research and all the Knapp's group, especially to my collaborator Florian Krull.

From left to right: Jan Zacharias, me and Anna Lena, Woelke.

Almost all the members in Knapp laboratory. From left to right: Tim Meyer, Prof. Knapp, me, Dawid Rasinski, Ilkay Sakalli, Anna Lena, Gerrit Korff, Jorge Numata and Jan Zacharias

References

[1] Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Proten-protein docking benchmark version 3.0. Proteins, 73: 705-709, 2008.
[2] Huang SY and Zou X. An iterative knowledge-based scoring function for protein-protein recognition. Proteins, 72: 557-579, 2008
[3] Pierce B, Weng Z. Structure Prediction of Protein Complexes. Proteins, 67: 1078-1086, 2007