行政院國家科學委員會專題研究計畫 成果報告

生物資訊在探索癌症相關基因上之研究--子計畫五：整合生物微晶片及蛋白質相互作用之資訊以改進與細胞週期有關之基因調控網路的預測(3/3)

研究成果報告(完整版)

計畫類別：整合型
計畫編號：NSC 95-2745-E-468-008-URD
執行期間：95年08月01日至96年07月31日
執行單位：亞洲大學生物科技學系

計畫主持人：吳家樂

計畫參與人員：碩士班研究生-兼任助理：蔡明成、寥晃聖、陳怡仲、邱金水
共同主持人：劉湘川

報告附件：出席國際會議研究心得報告及發表論文

處理方式：本計畫可公開查詢

中華民國 96年10月31日

私校能量計畫 - 子計畫五：
整合生物微晶片及蛋白質相互作用之資訊以改進與細胞週期有關之基因調控網路的預測
(3/3)–完整版報告

計畫主持人：吳家樂 **(Ka-Lok Ng)**

1. Introduction

The interaction between proteins is an important feature of protein function. Behind protein-protein interactions (PPI) there are protein domains interacting with each others to perform the necessary functions. Therefore, understanding proteins interactions at the domain level gives a global view of protein-protein interaction network (PIN). Putative domain-domain interactions (DDI) could be derived using the following approaches:

(1) association method (Deng et al. 2002),

(2) domain pair exclusion analysis (Riley et al. 2005),

(3) integrative approach (Ng et al. 2003a),

(4) domain combination pair approach, PreSPI (Han et al. 2004), and

(5) random decision forest model (Chen and Liu 2005).

2. Method

2.1 Input data

The domain combination pair approach (Han et al. 2004) is employed to derive putative protein DDI by using the PPI database DIP (Salwinski et al. 2004), Jan. 16, 2006 version, which recorded the PPI data for seven species: that is C. elegan, D. melanogaster, E. coli, H. pylori, H. sapiens, M. musculus and S. cerevisiae. Protein-domain annotation of DIP can be obtained from the protein domain database, Pfam (Finn et al. 2006). Pfam is a large collection of multiple sequence alignments for each domain family and uses hidden Markov models to find domains in new proteins. Domains in PfamA are well defined because the corresponding multiple sequence alignments and hidden Markov models have been checked, and most of the domains have been assigned functions.

2.2 Domain combination pair approach

Assuming a protein *A* contains *n* domains, there are 2^n-1 different domain combinations, the so-called power set of *A* with the empty set excluded, *ps'(A)*, according to the domain combination pair approach. Then given an interacting protein pair (*A*,*B*) with *m* and *n* domains respectively, one considers that there are (2^m-1)*(2^n-1) possible DDI. The set of domain combination pairs of two proteins A and B, *DC(A,B)*, is defined by

$$DC(A,B) = \{ps'(A) \times ps'(B)\} \qquad (1)$$

where × denotes the Cartesian product of set *ps'(A)* and *ps'(B)*. Since a protein can either has a single domain or multiple domains, combination of possible domain pairs can be derived from each of the interacting protein pair obtained from the DIP database (Salwinski et al. 2004).

To measure the likelihood of a DDI, the domain combination pair interaction matrix *M* is introduced. The element $M_{\alpha\beta}$ denotes the weighted interaction probability of a domain pair (α, β) for a given protein pair (A_i , B_j), and its value is given by

$$M_{\alpha\beta} = \sum_{(A_i,B_j)} \frac{1}{|ps'(A_i)| \times |ps'(B_j)|} \qquad (2)$$

where $|S|$ denotes the cardinality of set S, the summation is over all possible pairs of (A_i, B_j) such that α and β is an element of $ps'(A_i)$ and $ps'(B_j)$ respectively. Then, the elements of the normalized DDI interaction matrix $AP_{\alpha\beta}$ (so-called appearance probability matrix (Han et al. 2004) is defined by

$$AP_{\alpha\beta} = \frac{M_{\alpha\beta}}{\sum_{\alpha,\beta} M_{\alpha\beta}} \tag{3}$$

The matrix element $AP_{\alpha\beta}$ represents the DDI probability of domain combination α and β. If a DDI is found more frequently than expected by chance, it is likely that this DDI is a true interacting domain pair.

The results of putative DDI are computed for seven species. For each species, a negative learning set is constructed in order to improve the accuracy of DDI prediction. That is, given N proteins having K protein-protein interactions among them, the size of the negative learning set is equal to $C^N{}_2 + N - K$. This number represents the total number of non-interacting protein pairs for a particular species. Then, we calculated the probability of the DDI for the negative set of domain combination pairs, but now in the non-interaction space. Introduction of the negative learning set generated three AP matrices, one for the DDI space, I, one for the domain-domain non-interaction space (derived from the negative learning set), R, and the matrix elements of these two matrices are denoted by $AP^I{}_{\alpha\beta}$ and $AP^R{}_{\alpha\beta}$ respectively. The overlapping region of matrices AP^I and AP^R is denoted by AP^C, where C denotes the overlapping part. In other words, the domain combination pairs of two proteins A and B could be classified into three categories, that is $DC^I(A,B)$, $DC^R(A,B)$ and $DC^C(A,B)$.

After constructing the AP matrices, one can predict the interaction probability between the protein pair (A, B) based on the three AP matrices. Let X denotes the PPI and non-PPI events. A value of one and zero represent the PPI and non-PPI event respectively. Given the domain information for proteins A and B, one could determine the interaction probability using the Bayer's rule, that is

$$\begin{aligned} &P(X=1 \mid DC^C(A,B)) \\ &= \frac{P(X=1)\, P(DC^I(A,B) \mid X=1)}{P(X=1)\, P(DC^C(A,B) \mid X=1) + P(X=0)\, P(DC^C(A,B) \mid X=0)} \end{aligned} \tag{4}$$

where

$$P(X=1) \quad = \quad \frac{k \cdot I_{total} \cdot \sum_{\alpha,\beta} (AP_I^C)_{\alpha\beta}}{k \cdot I_{total} \cdot \sum_{\alpha,\beta} (AP_I^C)_{\alpha\beta} + (1-k) \cdot R_{total} \cdot \sum_{\alpha,\beta} (AP_R^C)_{\alpha\beta}} \tag{5}$$

where I_{total} and R_{total} in the above equations represent the total number of interacting and non-interacting protein pairs, respectively, $(AP_I^C)_{\alpha\beta}$ and $(AP_R^C)_{\alpha\beta}$ denote the interacting and non-interacting probability of domain combinations α and β in the overlapping space respectively, furthermore, $P(X=0) = 1 - P(X=1)$. The constant k is inserted into the Eq.(5) because the exact ratio of I_{total} and R_{total} in nature is unknown. The ratio of the total number of interacting and non-interacting protein pairs is determined by using the method of maximum-likelihood estimation. The maximal likelihood function L is defined by

$$L = C_x^n \; p^x (1-p)^{n-x} \tag{6}$$

where n is the total number of possible PPI, x is the total number of known PPI, and p is the probability of PPI. The parameter k is determined by the following condition,

$$\frac{\partial L}{\partial k} = 0 \qquad (7)$$

Once the probabilities of the non-interacting values for the domain combination pairs are obtained, then the probability of PPI is computed.

The probability that a protein pair (A, B) with m and n domains respectively could possibly interacting is estimated by the Primary Interaction Probability (PIP) (Han et al. 2004). PIP is given by

$$PIP(A-B) = \frac{\left\| AP^{I-C} \right\| + \left\| AP_I{}^C \right\|}{\left\| AP^{I-C} \right\| + \left\| AP_I{}^C \right\| + \left\| AP_R{}^C \right\|}(1 - P(X = 1 \mid DC^C(A,B)) \qquad (8)$$

where $AP^{I\text{-}C}$ denotes the matrix elements appear in the $AP^I - AP^C$ space, and $||AP||$ denotes the total sum of the matrix elements of AP.

In order to test whether the computed PIP results can provide potential PPI links between the proteins, three biological pathways (the yeast septin, E.coli chemotoxic pathway, and the blood coagulation pathway) are selected, then the pairwise PIP values for each pathway are computed and ranked, and the PPI prediction accuracy is determined by comparing with the corresponding experimentally determined network.

Three statistical measures are defined to characterize the prediction performance, that is the accuracy, Q, true positive specificity, S_{TP} , and true negative specificity, S_{TN}, they are defined as Q = (TP+TN)/(TP+TN+FP+FN), S_{TP} = TP/(TP+FP), and S_{TN} = TN/(TN+FP) respectively. TP, TN, FP and FN stand for true positive, true negative, false positive, and false negative events respectively.

2.3 Order index

Assuming that proteins A and B interacts, the AP-index of protein is defined by

$$H(A) = -\sum_{dA \in ps'(A)} p(dA) \log p(dA) \qquad (1)$$

where dA stands for an element of *ps'(A),* and $p(dA)$ denotes the DDI interaction probability of domain combination dA, For example, *H(A)* is replaced by $-\sum_{dA \in ps'(A)} AP^I_{dA} \log(AP^I_{dA})$. If *H(A)* is greater than *H(B),* then it is claimed that protein A regulates protein B. The rationale is based on the assumption that protein contains DDI with a larger AP^I value could possibly play the role of an upstream regulator.

3. Results

3.1 Putative DDI results and InterDom

To evaluate the prediction, the putative DDI results are compared with that of the database InterDom v.1.2 (June 2004). The results are depicted in Table 1.

In Table 1, we present the comparison of our putative DDI results with that of the database InterDom. All the DDI (N_{DDI}) are selected from our pre-computed DDI data, and compared with the InterDom records. Only DDI with a score larger than or equal to 0.4 are selected from InterDom in the comparison. InterDom assigns a score from 0 to 49322 for DDI, a score of 0.4 and above accounts for 90% of the 30037 records. The effective number of DDI, E_{DDI}, denotes single-domain interaction, and it does not include domain combination pair DDI, since these types of DDI are not available in InterDom. M_{DDI} stands for the matched DDI, and the matching ratio S_M is defined as M_{DDI}/E_{DDI} *100%.

Table 1. The putative domain-domain interaction results (N_{DDI}) obtained by the domain combination pair approach compared with that of the database InterDom.

In the InterDom comparison study, the DDI matching ratio ranges from 66.3% to 89.5% for the seven species. An average matching ratio of 75.7% is obtained, this indicates the model is rather sucessiveful.

3.2 The yeast septin complex

To verify whether the pre-computed DDI results can provide potential PPI links between proteins, three biological pathways, i.e. the septin complex, the E. coli chemotaxis pathway and the blood coagulation pathway, are selected for further study. The predicted PPI events among those proteins in these three pathways are compared with the experimentally determined PPI networks.

For the first comparison, the yeast septin complex, which composed of six proteins (CDC3, CDC10, CDC11, CDC12, GIN4, SHS1), is selected. Thus, there are 15 (C^6_2) possible PPI among the proteins. A PPI link is assumed if the PIP value is larger than or equal to 0.1, our method correctly predicted that all six proteins interact with each other, that is a prediction accuracy Q of 100% (15/15) is achieved as well as S_{TP} and S_{TN}. The same prediction accuracy is reported by InterDom. In contrast, PreSPI returned an error message (PIP_Value = error) for each of the 15 possible interactions. In this case, our prediction performed much better than PreSPI.

The PIP threshold is set at the 0.1 level because this is the least stringent value among the three PPI cases we studied. Use of a small PIP threshold would predict more PPI, but most of them are false positive interactions. In order to show how the threshold affects the prediction accuracy and specificity performance, a higher threshold value of 0.6 is selected for further study, and the results are reported in section 3.3 and 3.4

3.3 The E. coli chemotaxis pathway

In the second study, the E. coli chemotaxis pathway is selected. Chemotaxis is the response of cells to chemical stimuli by directed movement. The chemotaxis pathway, obtained from KEGG (Kanehisa et al. 2006), composed of 11 proteins: MCP (consists of trg, tap, CheD, CheM), Aer, CheA, CheB, CheR, CheW, CheY, and CheZ. The predicted results based on DDI are depicted in Table 2 (a PPI is assumed if the PIP threshold is set to 0.1 or 0.6).

Table 2. Predicted number of protein-protein interactions and the statistical measure results with PIP threshold of 0.1 and 0.6 in the *E. coli* chemotaxis pathway, and compared with that of PreSPI.

For the chemotaxis pathway, there are 55 (C^{11}_2) possible interactions among the 11 proteins. Our prediction returned a PIP value for each of the 55 interactions. PreSPI returned only 36 interactions, and the rest are not addressed. The accuracy of our prediction is comparable (at the 0.1 threshold level) to PreSPI, whereas 19 more PPI links are predicted, and a better true negative specificity S_{TN} are obtained. The InterDom database gave null result for this pathway study. If the threshold is set to the 0.6 level, it gave a much better sensitivity and specificity ratios, for instance, the accuracy, Q, raised from 51% to 76%, the specificity ratios, S_{TP} and S_{TN}, raised from 33% to 57%, and 37% to 85%, respectively.

3.4 The blood coagulation pathway

In the last study, we applied the computed DDI data to reconstruct the blood coagulation pathway. Blood clotting occurs via three pathways, intrinsic, extrinsic and common pathways, in which a total of 13 proteins are involved. The blood coagulation pathway composed of 13 proteins: FI, FII, FIII, FV, FVII, FVIII, FIX, FX, FXI, FXII, FXIII, PKK, and HMWK. Based on the DDI data, the predicted results are depicted in the Table 3. In general there are 78 possible interactions, but only 48 interactions can be determined in our computation (a PPI is assumed if the PIP threshold is set to 0.1 or 0.6)

Table 3. Predicted number of protein-protein interactions and the statistical measure results with PIP threshold of 0.1 and 0.6 in the blood coagulation pathway, and compared with that of PreSPI.

When comparing our results with those predicted by PreSPI, our prediction achieves a much better accuracy (at the 0.1 threshold level). Both computations returned similar S_{TP} value, however, our calculation obtained a much better value of S_{TN}. If the threshold is set to the 0.6 level, it gave a slightly better sensitivity and specificity ratios, for instance, the accuracy, Q, raised from 54% to 60%, the specificity ratios, S_{TP} and S_{TN}, raised from 24% to 28%, and 57% to 65%, respectively.

The difference between our results and that of PreSPI is probably because of PreSPI used the IntAct (Hermjakob et al. 2004) database for domain annotations, whereas the Pfam database is used in our work. It is known that the two databases provide a somewhat different set of domain annotations for proteins, this leads to the fact that different inputs (the learning set as well as the negative learning set) are used by each study.

To further characterize PPI, the regulatory orders of PPI for six biological pathways are studied, and the results are given in the following sub-sections. All the pathways are taken from E.coli and yeast only, since the PPI data and domain annotations coverage rate for these two species are relative higher than the other five species, in other words, the problem of missing domain annotations and DDI information are less severe in those two species.

3.5 Order index - E. coli chemotaxis pathway

The chemotaxis pathway composed of six proteins or protein complexes: MCP, Aer, (CheA,CheW), CheB, CheY, and CheZ. The following five PPI regulatory order pairs are recorded in KEGG: MCP-(CheA-CheW), Aer-(CheA,CheW), (CheA,CheW)-CheB, (CheA,CheW)-CheY, and CheZ-CheY, where the bracket (…..) stands for protein complex, and symbol on the left of a regulatory relation X-Y is the upstream regulatory protein.

For the chemotaxis pathway, the *AP*-order index approach correctly predicted the five regulatory relationships, it achieves a prediction accuracy of 100% (i.e. 5/5).

The same method is applied for the other five PPI pathways as well. It is demonstrated in the following subsections that the prediction accuracy of the order index approach is very encouraging.

3.6 Order index – the yeast cell cycle DNA damage checkpoint

The yeast cell cycle DNA damage checkpoint in the G2 phase is selected in this study. In this pathway there are 20 regulatory relations among the following 22 proteins or protein complexes: Rad17, Rad24, Mec3, Ddc1, Rad9, Mec1, Ctr1, Chk1, Pds1, Rad53, Cdc5, (Clb1, Cdc28), Mih1, Cak1, Cks1, Swi5, Sic1, Swe1, (Scf, Met30), Gin4, Hsl1, (Hsl7, Hsl1). Since the domain annotation for Mec3, Ddc1, Pds1 and Sic1 are not available (four PPI relations are removed), the regulatory relations for Rad17-Rad24, and (Clb1, Cdc28)-Cks1 are not clearly defined by KEGG (two more regulatory relations are removed), therefore, only 14 relations (the second column in Table 4) among 15 proteins are considered in the prediction.

Among the 14 relations, the relative dependence of Gin4-Swe1 and Hsc1-Swe1 are not determined because they have the same *AP*-order index values, hence, only 12 relations left (the third column in Table 4). Among the 12 relations, 7 relations are correctly predicted. The seven correct predictions are: Rad9-Mec1, Mec1-Chk1, Mec1-Rad53, Cdc5-(Clb1,Cdc28), Mih1-(Clb1,Cdc28), Cak1-(Clb1, Cdc28) and Swe1-(Clb1,Cdc28). Hence, the regulatory order prediction accuracy for the damage checkpoint pathway is 58.3% (i.e. 7/12).

3.7 Order index – the yeast cell cycle spindle checkpoint

For the spindle checkpoint pathway, there are 12 regulatory relations among the following 14 proteins or protein complexes: Mps1, (Bub1,Bub3), (Mad1,Mad2,Mad3), (APC/C, Cdc20), (APC/C, Cdh1), Cdc14, Swi5, Sic1, Esc5, (Dbf2, Mob1), Dbf20, Tem1, Bub2 and Let1. Since the domain annotation for Sic1 and Esc5 are not recorded in the SwissProt database, therefore, three of the protein regulatory relations cannot be determined. Furthermore, one relation has the same *AP*-order index value ((APC/C, Cdc20)-(APC/C, Cdh1)), hence 8 relations left. The order index method correctly predicted seven PPI regulatory order pairs out of the eight relations, these are Mps1-(Bub1,Bub3), (Mad1,Mad2,Mad3)-(APC/C,Cdc20), Cdc14-(APC/C, Cdh1), Let1-Tem1, Tem1-Dbf20, Cdc14-Swi5, and Bub2-Tem1. Hence, the method achieves a prediction accuracy of 87.5% (i.e. 7/8).

3.8 Order index – the yeast MAPK signaling pathway, starvation

In this study the yeast starvation, osmolarity and hypotonic shock pathways are selected. For the starvation pathway, there are six regulatory relations among the following seven proteins, Sho1, Ras2, Cdc42, Ste20, Ste11, Ste7 and Kss1. Among the six relations, the relative dependence of Ste7-Kss1 is not determined because it has the same *AP*-order index value. The order index method correctly predicted the regulatory order of the other five PPI pairs: Ras2-cdc42, Sho1-Cdc42, Cdc42-Ste20, ste20-ste11, ste11-ste7, ste7-Kss1. The method achieves a prediction accuracy of 100% (5/5).

3.9 Order index – the yeast MAPK signaling pathway, osmolarity

For the osmolarity pathway, there are eight regulatory relations among the following nine proteins: Sho1, Sln1, Ste20, Ypd1, Ste11, Ssk1, Ssk2, Pbs2, and Hog1. Among the eight relations, the PBs2-Hog1 relation has the same *AP*-order index value, hence seven relations left. The order index method correctly predicted the six PPI relations: Sho1-Ste20, ste20-ste11, ste11-Pbs2, Ypd1-Ssk1, Ssk1-Ssk2, and Ssk2-Pbs2, hence, the method achieves a prediction accuracy of 85.7% (6/7)

3.10 Order index – the yeast MAPK signaling pathway, hypotonic shock

For the hypotonic shock pathway, there are six regulatory relations among the following seven proteins: Mid2, Rho1, Fks1, Pkc1, Bck1, (Mkk1,Mkk2), and Slt2. Among the six relations, two relations have the same *AP*-order index values, hence, four relations left. The order index method correctly predicted the three PPI relations: Mid2-Rho1, Fks1-Rho1 and Rho1-Pkc1, and Pkc1-Bck1 is incorrectly predicted, hence, the method achieves a prediction accuracy of 75% (3/4).

In Table 4, we summarized the total number of PPI relations recorded by KEGG, the total number of PPI with well-defined domain annotation, the number of correct predictions determined by the order index method, and the prediction accuracy for the six pathways we selected. On average the order index approach can achieved a prediction accuracy of 80.5%, that is, for the six PPI pathways we studied, 33 relations are correctly predicted among a total of 41 relations. A total of 48 PPI relations are studied, in which seven relations have the same *AP*-order index values, hence, the coverage rate of prediction is 85.4%.

Table 4. The prediction accuracy of the *AP*-order index method with the threshold set to 1.0. The first column denotes the name of the studied pathway. The second column represents the total number of PPI relations recorded in KEGG with which domain annotation are well-defined. The third column represents the number of PPI relations left after taking into account of the threshold. The fourth column represents the number of regulatory orders correctly predicted by the *AP*-order index method. The last column denotes the prediction accuracy of the method.

3.11 Order index – robustness test

In order to test the robustness of the order index calculation, we assumed that if the *AP*-order index values for two regulatory relations differed by least than 10% (the difference between the larger value and the smaller value divided by the smaller one), then we considered that the method is not able to determine the regulatory order. The regulatory order predictions are repeated for the above six pathways, and the results is depicted in Table 5. The order index approach predicted 25 correct relations out of 31 relations, this amounts to a prediction accuracy of 80.6%, which is essentially the same as the prediction without the 10% difference choice. This indicates that the order index approach is rather robust with respect to the choice of threshold. The coverage rate of regulatory order prediction is equal to 64.6%, i.e. 31/48.

Table 5. The prediction accuracy of the *AP*-order index method with the threshold set to 1.1. The first column denotes the name of the studied pathway. The second column represents the total number of PPI relations recorded in KEGG with which domain annotation are well-defined. The third column represents the number of PPI relations left after taking into account of the threshold. The fourth column represents the number of regulatory orders correctly predicted by the *AP*-order index method. The last column denotes the prediction accuracy of the method.

To account for the statistical significance of the method, a hypothesis test is performed on the mean number of correct predictions for the six pathways. Assuming a one-tailed binomial probability distribution test, the hypothesis *t*-test rejects the null hypothesis at a 99% level.

3.12 A web-based service for PPI and regulatory order prediction

The predicted domain-domain interaction results are available at http://210.70.82.82/kzbio2/r_ap.php. Several query interfaces are implemented to facilitate data display, such as the DDI, PIP, PIP query and network reconstruction services. For instance, the PIP query service allows the user to input two proteins' Swissprot (Boeckmann et al. 2003) ID and get the probability of their interaction, i.e. the PIP value. In case the actual Swissprot ID is not known, user can input domain's PfamA ID, the system could return the predicted probability of the protein interaction. Furthermore, the network service allows the user to reconstruct PPI network, and predict the regulatory order of a PPI. To reconstruct the PPI network, user can either input a set of proteins or domains IDs, the system returns a text file where putative PPI interactions are predicted. The predicted PPI network can be visualized by reading the output file using Cytoscape (Shannon *et al.* 2003).

We also have set up a web-based service for the public to use the *AP*-order index method for prediction, which is available at http://210.70.82.82/kzbio2/oi.php. For instance, if one wants to determine the regulatory order of Aer and (CheA, CheW), prepare the following line as an input,

```
PF08447,PF00672,PF00015    PF01627,PF02895,PF02518,PF01584
```

where the first and second columns denote the PfamA annotations of the Aer and (CheA, CheW) proteins.

Paste the above line in the box provided in the AP-order index web page, give a name for the output file, select *E.coil* under the species manual, and press the send button. The platform will return a file which states the prediction result (either A regulates B or not able to determine the regulatory order).

4. Conclusion

The domain combination pair approach is employed to derive putative protein DDI from the PPI database DIP. To evaluate the prediction performance of the approach, the DDI predicted results are compared with that of the database InterDom, where an average matching ratio of 75.7% can be achieved (assuming the Jan. 16, 2006 version of DIP).

Three PPI networks are chosen to test the prediction accuracy of our computation. The yeast septin complex, and the blood coagulation pathways are reconstructed with a much better accuracy and true negative specificity than another study. For the E. coli chemotaxis pathway study, comparable PPI prediction accuracy is obtained whereas more PPI and a better true negative specificity are obtained in our prediction. This indicated the merit of our calculations. Furthermore, an entropy-like quantity, so called *AP*-order index, based on DDI data, is introduced to predict the regulatory order for a PPI. The prediction accuracy of this method is demonstrated for six PPI pathways. It is found that this method can achieve a prediction accuracy of 80.5%. This implies that the order index is a very reliable parameter to determine the regulatory order of PPI.

There are two major obstacles for the PPI and regulatory order calculations: (i) many proteins do not have complete PfamA domain annotations, and (ii) there is the missing DDI information problem. Much further experimental works are still needed to resolve prior two problems.

Acknowledgements

Drs. Ka-Lok Ng and Chien-Hung Huang works are supported by the National Science Council of R.O.C. under the grant of NSC 95-2745-E-468-008-URD and NSC 95-2221-E-150-025-MY2 respectively. We thank Mr. Jengru Lee for developing computer programs to do the calculation, and constructing the database.

References

Boeckmann B., Bairoch A., Apweiler R., Blatter M.C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., and Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res., 31, 365 - 370.

Chen Xue-Wen and Liu Mei 2005. Prediction of protein-protein interactions using random decision forest framework. Bioinformatics 21, 4394-4400.

Deng Minghua, Mehta Shipra, Sun Fengzhu, and Chen Ting 2002. Inferring domain-domain interactions from protein-protein interactions. Genome Res. 12, 1540-1548.

Finn R. D., Mistry Jaina, Schuster-Böckler Benjamin, Griffiths-Jones Sam, Hollich Volker, Lassmann Timo, Moxon Simon, Marshall Mhairi, Khanna Ajay, Durbin Richard, Eddy Sean R., Sonnhammer Erik L. L., and Bateman Alex 2006. Pfam: clans, web tools and services. Nucl. Acids Res. 34, D247-D251.

Han Dong-Soo, Kim Hong-Soog, Jang Woo-Hyuk, Lee Sung-Doke, and Suh Jung-Keun 2004. PreSPI: a domain combination based prediction system for protein-protein interaction Nucleci Acids Res. 32, 6312-6320.

Hermjakob H., Montecchi-Palazzi L., Lewington C., Mudali S., Kerrien S., Orchard S., Vingron M., Roechert B., Roepstorff P., Valencia A., Margalit H., Armstrong J., Bairoch A., Cesareni G., Sherman D., Apweiler R. 2004. Nucl. Acids. Res. 32: D452-D455.

Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K.F., Itoh M., Kawashima S., Katayama T., Araki M., and Hirakawa M. 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354-357.

Ng S.K., Zhang Z., Tan S.H. 2003a. Integrative approach for computationally inferring protein domain interactions. Bioinformatics 19, 923-929.

Ng S.K., Zhang Z., Tan S.H., Lin K. 2003b. InterDom: a database of putative interaction protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res., 31, 251-254.

Riley R., Lee C., Sabatti C., Eisenberg D. 2005. Method Inferring protein domain interactions from databases of interacting proteins. Genome Biology 6, R89.

Salwinski L., Miller C.S., Smith A.J., Pettit F.K., Bowie J.U., Eisenberg D. 2004. The Database of Interacting Proteins. Nucl. Acids Res. 32, D449-51.

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498-504

Table 1. The putative domain-domain interaction results (N_{DDI}) obtained by the domain combination pair approach compared with that of the database InterDom.

Species	N_{DDI}	E_{DDI}	M_{DDI}	S_M(%)
C. elegan	3874	1751	1142	65.2
D melagonster	1346	695	523	75.4
E .coli	59062	12075	1695	66.3
H. pylori	894	276	247	89.5
H. sapiens	6327	1187	849	71.5
M. musculus	1031	206	160	77.7
S. cerevisiae	39415	4440	3772	84.5
Average				75.7%

Table 2. Predicted number of protein-protein interactions and the statistical measure results with PIP threshold of 0.1 and 0.6 in the *E. coli* chemotaxis pathway, and compared with that of PreSPI.

	Threshold = 0.1	Threshold = 0.6	PreSPI
TP	13	8	12
TN	15	34	7
FP	26	6	17
FN	1	7	0
total	55	55	36
Q	51%	76%	53%
S_{TP}	33%	57%	41%
S_{TN}	37%	85%	29%

Q = (TP+TN)/(TP+TN+FP+FN), S_{TP} = TP/(TP+FP), S_{TN} = TN/(TN+FP), total = TP+TN+FP+FN.

Table 3. Predicted number of protein-protein interactions and the statistical measure results with PIP threshold of 0.1 and 0.6 in the blood coagulation pathway, and compared with that of PreSPI.

	Threshold = 0.1	Threshold = 0.6	PreSPI
TP	5	5	18
TN	21	24	0
FP	16	13	60
FN	6	6	0
total	48	48	78
Q	54%	60%	23%
S_{TP}	24%	28%	23%
S_{TN}	57%	65%	0%

Q = (TP+TN)/(TP+TN+FP+FN), S_{TP} = TP/(TP+FP), S_{TN} = TN/(TN+FP), total = TP+TN+FP+FN.

Table 4. The prediction accuracy of the *AP*-order index method with the threshold set to 1.0. The first column denotes the name of the studied pathway. The second column represents the total number of PPI relations recorded in KEGG with which domain annotation are well-defined. The third column represents the number of PPI relations left after taking into account of the threshold. The fourth column represents the number of regulatory orders correctly predicted by the *AP*-order index method. The last column denotes the prediction accuracy of the method.

Pathway name	Total no. of PPI relations	Actual no. of PPI relations	Correct predictions	Accuracy (%)
Chemotaxis	5	5	5	100
DNA damage	14	12	7	58.3
Spindle checkpoint	9	8	7	87.5
Starvation	6	5	5	100
Osmolarity	8	7	6	85.7
Hypotonic	6	4	3	75.0
Total	48	41	33	

Table 5. The prediction accuracy of the *AP*-order index method with the threshold set to 1.1. The first column denotes the name of the studied pathway. The second column represents the total number of PPI relations recorded in KEGG with which domain annotation are well-defined. The third column represents the number of PPI relations left after taking into account of the threshold. The fourth column represents the number of regulatory orders correctly predicted by the *AP*-order index method. The last column denotes the prediction accuracy of the method.

Pathway name	Total no. of PPI relations	Actual no. of PPI relations	Correct predictions	Accuracy (%)
Chemotaxis	5	3	3	100
DNA damage	14	10	7	70.0
Spindle checkpoint	9	8	7	87.5
Starvation	6	2	2	100
Osmolarity	8	4	3	75.0
Hypotonic	6	4	3	75.0
Total	48	31	25	

Self-assessment

We have completed the major aims of the proposal, that is deriving putative domain-domain interaction pair and introduce the AP-index to predict the regulatory order of a protein-protein interaction pair. A web-based service was set up which provide the PPI and regulatory order services for the public.

During the period 2004 and 2005, our results are presented, either oral or poster presentations, in international conferences and local conferences.

Publications

期刊論文

1. Ka-Lok Ng, Chien-Hung Huang*, Hsueh-Chuan Liu, Hsiang-Chuan Liu (2008)
   Applications of domain-domain interactions in pathway study
   ***Computational Biology and Chemistry,* 32** (in press, to appear at 2008)
2. J.D. Wang, Hsiang-Chuan Liu, Jeffrey J.P. Tsai, **Ka-Lok Ng*** (2007)
   Scaling Behavior of Maximal Repeat Distributions in Genomic Sequences
   Int'l J. of Cognitive Informatics and Natural Intelligence (to appear)
3. Kuo-Ching Hsiao, Chien-Hung Huang, **Ka-Lok Ng*** (2006)
   Protein Structural Classes Prediction via Residues Environment Profile
   Asian J. Health and Information Sci., 1(3), in press
4. Chien-Hung Huang, Jywe-Fei Fang, Jeffrey J.P. Tsai, **Ka-Lok Ng*** (2007)
   "Topological Robustness of the Protein-protein Interaction Networks"
   ***Lecture Notes in Bioinformatics* vol. 4023**, **RECOMB 2005 Regulatory Genomics and Systems Biology Workshop,** E. Eskin et al . (Eds.), p.166-177, Springer Verlag (SCI 著作)

* corresponding author

國際性研討會論文或壁報

2007 年

1. Liu Hsueh-Chuan, Huang Chien-Hung, Tsai J.F, **Ng Ka-Lok*** "APPLICATIONS OF DOMAIN-DOMAIN INTERACTION IN PATHWAYS STUDY" 5th Asia-Pacific Bioinformatics Conference (**APBC2007**) Hong Kong, 15-17, Jan. 2007, poster abstract p.42. **(**95 學年)

國際性研討會論文或壁報

2006 年

1. Lee Jeng-ru, Liu Hsiang-Chuan, Tsai J.F., **Ng Ka-Lok***. "Large scale prediction of domain-domain interactions from protein-protein interactions". 4th Asia-Pacific Bioinformatics Conference (**APBC2006**) Taiwan, 13-16 Feb, 2006. P063 **Poster (**94 學年)
2. Huang Chien-Hung, Tsai J.F., Fang Jywe-Fei, **Ng Ka-Lok***. "Topological Stability of the protein-protein interaction networks". 4th Asia-Pacific Bioinformatics Conference (**APBC2006**) Taiwan, 13-16 Feb, 2006. P062 **Poster (**94 學年)
3. Wang J.D., Liu Hsiang-Chuan, **Ng Ka-Lok***. "Scaling Behavior of Maximal Repeat Distributions in Genome Sequences". 4th Asia-Pacific Bioinformatics Conference (**APBC2006**) Taiwan, 13-16 Feb, 2006. P061 **Poster (**94 學年)
4. Chien-Hung Huang, Tsai J.F., **Ng Ka-Lok***. "Deriving Domain-domain Interactions from Protein-protein

Interactions Networks”. **INFORMS06**, Hong Kong, 25-28, June 2006, p.40. **Oral presentation (**94 學年)

5. **Ng Ka-Lok***, Liu Hsueh-Chuan, Liu Hsiang-Chuan, Tsai J.F. “Reconstructing protein-protein interaction networks from domain-domain interactions”. Asia Pacific Association for Medical Informatics (**APAMI 2006), Taipei,** October 27-29, 2006**. p.31, Oral presentation 全文 (**95 學年)

6. Hsiang-Chuan Liu, Chien-Hung Huang, **Ka-Lok Ng** “Protein-protein interaction pathways reconstruction from domain-domain interactions”. **The 7th International Conference on Systems Biology** (ICSB-2006), Yokohama Japan, 9-11 October 2006. **Poster, p.42, FN43 (**95 學年)

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

95 年 10 月 16 日

附件三

報告人姓名	吳家樂	服務機構及職稱	亞洲大學 生物科技與生物資訊系 副教授
時間 會議 地點	8 – 12 Oct. 2006 ICSB 2006 Pacifico Yokohama, Japan	本會核定補助文號	NSC 95-2745-E-468-008-URD
會議名稱	(中文)第七屆系統生物學 2006 國際研討會 (英文) The 7th International Conference on System Biology 2006		
發表論文題目	(中文)從蛋白質功能域相互作用推測蛋白質相互作用網路 (英文)Protein-protein interaction pathways reconstruction from domain-domain interactions (中文)人類 miRNA 基因與調控子及 CpG 島 (英文)Finding human miRNA genes located within promoter regions and associated with CpG islands		

報告內容應包括下列各項：

一、參加會議經過

Oct. 8

- Morning session : Attended Tutorial 8 - Modeling, simulating, and analyzing biochemical systems with Copasi
- Afternoon session : Attended Tutorial 6 - Analyzing Biochemical Systems using the E-Cell System

Oct. 9

- Attended the Plenary Talks: P1 , P2, P4: Oct. 9th 10:00-12:30
- P1: Upinder S. Bhalla (The National Centre of Biological Science, Bangalore)
  "Electricity meets Chemistry: Fast and Slow Signaling in Memory "
- P2: Atsushi Miyawaki (Riken Brain Science Institute)
  "Spatio-temporal Patterns of Intracellular Signaling"
- P4: Luis Serrano (European Molecular Biology Laboratory)
  "Evolvability and hierarchy in rewired bacterial gene networks"

Oct. 10

- Attended the Plenary Talks: P3, P5: Oct.10, 9:00-9:30, 9:30-10:00
- P3: Stephen Quake (Stanford University / HHMI)
  "Biological Large Scale Integration"
- P5: Steve Oliver (University of Manchester)
  "Dealing with the complexity of a 'simple' eukaryotic cell"

Oct. 11

- **Complex Systems Biology - Oct. 11th, 9:15-10:30**
- Mihajlo D. Masarovic (Case Western Reserve University)
  "Interaction Balance Coordination as Organizing Principle in Complex Systems Biology"
- Jack Donald Keene (Duke University Medical Center)
  "Coordination of Gene Expression by RNA Operons"
- Kenneth Alan Loparo (Case Western Research University)
  "Applications of Complex Systems Biology to the Study of Neural Systems "
- **Control and System Theory for Systems Biology - Oct. 11th, 11:00-12:30**
- Francis J. Doyle (University of California, Santa Barbara)
  "Robustness Analysis of Biological Networks Using Sensitivity Measures"
- Pablo A. Iglesias (Johns Hopkins University)
  "Feedback Control Regulation of Cell Division"
- John Doyle (California Institute of Technology)
  "The architecture of cellular regulation"
- **Signal transduction - Oct. 11th, 14:00-16:30**
- Hans V. Westerhoff (The University of Manchester)
  "Cell-signaling Dynamics in Time and Space"
- William S. Hlavacek (Los Alamos National Laboratory)
  "Rules for Modeling Signal-Transduction Systems"
- Philippe Bastiaens (EMBL Heidelberg)
  "Reaction cycles in the spatial and temporal organization of cell signaling"

二、與會心得

This year the conference was held at Yokohama, Japan. The conference comprises a broad area of topics in the area of system biology.

Topics include,

- Systems Biology for Medicine
  Drug discovery, Cancer Systems Biology, Systems Immunology, Cardiovascular Systems

Biology, Systems Biology of Diabetes and Metabolic Syndrome

- Systems Biology of Basic Biological Processes
  Developmental Systems Biology, Metabolome and Bioprocess, Signal transduction, Cyclic and Dynamical Behaviors, Systems Neuroscience, Microorganisms

- Expanding Fronts in Systems Biology
  Large-Scale Biology, Bioinformatics Support for Systems Biology, Synthetic Biology, Complex Systems Biology, Systems marine biology
  I had two posters presentation on Oct. 9 and 10. The titles of my two posters presentations were (i) Protein-protein interaction pathways reconstruction from domain-domain interactions and, (ii) Finding human miRNA genes located within promoter regions and associated with CpG islands.

There were many interesting and good talks presented in this conference. I highlighted their main interesting results in below.

Oct. 8 9:30 am
Tutotial - Modeling, simulating, and analyzing biochemical systems with Copasi
Pedro Mendes (Virginia Bioinformatics Institute)

Copasi (Complex Pathway Simulator) is a software application for simulation and analysis of biochemical networks. It is developed jointly by the groups of Pedro Mendes (Virginia Bioinformatics Institute, USA) and Ursula Kummer (EML Research, Germany), and is freely available for academic use.

Copasi's current features include stochastic and deterministic time course simulation, steady-state analysis (including stability), metabolic control analysis, elementary mode analysis, mass conservation analysis, import and export of SBML level 2, optimization, parameter scanning and parameter fitting. It runs on MS Windows, Linux, OS X, and Solaris SPARC. So, it is one of the few computational tools in systems biology that are OS X compatible.

The presenters use Copasi to explain how the modelling, simulation and computational analysis of biochemical systems works. They also critically evaluate the limitations of different simulation methods.

Oct. 8 1:30 pm
Tutotial - Analyzing Biochemical Systems using the E-Cell System
Nathan Addy, Satya Arjunan, Bin Hu, Yuri Matsuzaki, Martin Robert, Takeshi Sakurada Koichi Takahashi (Keio University)

Bifurcation and sensitivity analysis can be used to elucidate the relationship between the dynamics of a nonlinear system in biology and the parameters of the system. The bifurcation program in E-Cell numerically computes the stable states of the system, such as the stable or oscillating point, with graphical representation of results. Elasticity coefficients with respect to amplitude and frequency, which indicate the robustness of the oscillation are also represented. Participants will experiment with these features hands-on using a simple oscillation model – the Drosophila circadian cycle model.

Metabolic control analysis can demonstrate how fluxes and intermediate concentrations in a metabolic pathway are regulated by the enzymes that constitute the system. The analysis encompasses structural analysis, elasticity coefficients and the sensitivity of metabolites to small changes in individual parameters such as in enzyme concentrations or kinetic parameters. Flux and concentration control coefficients are some of the outcomes of metabolic control analysis. Participants used metabolic control analysis to evaluate the Kuchel's erythrocyte model.

October 9 10:00-10:30 am

“Electricity meets Chemistry: Fast and Slow Signaling in Memory”

Upinder S. Bhalla

National Centre for Biological Sciences, TIFR, Bangalore, India
http://www.ncbs.res.in/~bhalla/index.html

Deliberations on memory mechanisms often seem to proceed on at least three independent tracks. One of these involves biochemical mechanisms for plasticity, including feedback loops and cellular activation. Space is another dimension, and is the arena for interactions between synapses, and propagation of signals between synapses, dendrites, and the cell body. Finally, electrical activity is a function of cell as well as network dynamics, and here too feedback may play a role through reverberating activity in network loops. It is an interesting process to develop models that impinge on all of these levels, because of the wide range of timescales, numerical techniques, and sheer computational load. It is especially tricky to get parameters for such models. I will describe a study where we have used coupled electrical and biochemical compartmental modeling, and weeded out several candidate models by comparing their predictions to our experiments. The surviving models incorporate chemical, spatial and electrical ingredients. They exhibit network-activity controlled single-cell reverberating activation, with interesting spatial consequences. We suggest that this is a form of short-term and spatially defined memory. It sits at the interface between individual synapses and dendrites, and also between network and cellular attributes of memory.

October 9 10:30-11:00 am

“Spatio-temporal Patterns of Intracellular Signaling”

Atsushi Miyawaki

RIKEN Brain Science Institute
http://www.brain.riken.go.jp/english/b_rear/b5_lob/a_miyawaki.html

“Why bio-imaging, i.e. real time fluorescence imaging?" Currently, this is a topic of great interest in the bioscience community. Many molecules involved in signal transduction have been identified, and the hierarchy among those molecules has also been elucidated. It is not uncommon to see a signal transduction diagram in which arrows are used to link molecules to show enzyme reactions and intermolecular interactions. To obtain a further understanding of a signal transduction system, however, the diagram must contain the three axes in space as well as a fourth dimension, time, because all events are controlled ingeniously in space and time. Since the isolation of green fluorescent protein (GFP) from the bioluminescent jellyfish in 1992 and later with its relatives, researchers have been awaiting the development of a tool, which enables the direct visualization of biological functions. This has been increasingly enhanced by the marriage of GFP with fluorescence resonance energy transfer (FRET) or fluorescence cross-correlation spectroscopy (FCCS), and is further expanded upon by the need for "post-genomic analyses." It is not my intent to discourage the trend seeking the visualization of biological function. I would like to propose that it is time to evaluate the true asset of "bio-imaging" for its potential and limitations in order to utilize and truly benefit from this novel technique.

October 9 12:00-12:30 pm

"Evolvability and hierarchy in rewired bacterial gene networks"

Luis Serrano

EMBL-CRG Systems Biology Programme, Centre for Genomic Regulation, Spain, 2. EMBL, Germany
http://www-db.embl.de/jss/EmblGroupsHD/per_397.html

Bacterial gene networks are highly plastic, allowing radical reconnections at the summit of the gene network hierarchy, fuelling evolvability.Sequencing of genetic material from several organisms has revealed that duplication and drift of existing genes has primarily molded the contents of a given genome. Though the effect of knocking out or over-expressing a particular gene has been studied in many organisms, no study has systematically explored the effect of adding new links in a biological network. To explore network plasticity, we constructed 598 recombinations of promoters (including regulatory regions) with different transcription or s-factors in Escherichia coli, over the genetic background of the wild-type. We found that ~95% of reconnected networks are tolerated by the bacterial cell and very few give different growth profiles. Expression levels correlate with the position of the factor in the wild-type network hierarchy. Most importantly, we find that certain combinations consistently survive over the wild-type under various selection pressures. This suggests that new links in the network could readily confer a fitness advantage to individuals in a population and hence may fuel evolution.

October 10 9:00 – 9:30 am

"Biological Large Scale Integration"

Stephen Quake

Dept of Bioengineering and (by courtesy) Applied Physics, Stanford University and Howard Hughes Medical Institute

http://med.stanford.edu/profiles/Stephen_Quake/

The integrated circuit revolution changed our lives by automating computational tasks on a grand scale. My group has been asking whether a similar revolution could be enabled by automating biological tasks. To that end, we have developed a method of fabricating very small plumbing devices – chips with small channels and valves that manipulate fluids containing biological molecules and cells, instead of the more familiar chips with wires and transistors that manipulate electrons. Using this technology, we have fabricated chips that have thousands of valves in an area of one square inch. We are using these chips in applications ranging from bioreactors to structural genomics to systems biology. However, there is also a substantial amount of basic physics to explore with these systems – the properties of fluids change dramatically as the working volume is scaled from milliliters to nanoliters.

- Microfluid system
- Large half-life of protein function
- Biological dark matter – 99% of bacteria cannot be cultivate

October 10 9:30-10:00 am

"Dealing with the complexity of a 'simple' eukaryotic cell"

Stephen G. Oliver

Faculty of Life Sciences, The University of Manchester, U.K.

http://www.ls.manchester.ac.uk/people/profile/index.asp?tb=0

Systems biology aims at taking a more synthetic or holistic approach to deciphering the workings of living organisms. Although the ultimate aim is to construct mathematical models of complete cells or organisms that have both explanatory and predictive power, we are some way from achieving such global syntheses and we need a principled way of reducing the complexity of the problem. Accordingly, we require a top-down strategy to provide an initial coarse-grained model of the cell, and a bottom-up strategy in which individual sub-systems are modeled.

Metabolic Control Analysis (MCA) is a conceptual and mathematical formalism that models the relative contributions of individual effectors in a pathway to both the flux through the pathway and the concentrations of individual intermediates within it. To exploit MCA in an initial top-down systems analysis of the eukaryotic cell, two categories of experiments are required. In category 1 experiments, flux is changed and the impact on the levels of the direct and indirect products of gene action is measured. We have measured the impact of changing the flux on the transcriptome, proteome, and metabolome of Saccharomyces cerevisiae. In this whole-cell analysis, flux equates to growth rate. In category 2 experiments, the levels of individual gene products are altered, and the impact on the flux is measured. We have used competition analyses between the complete set of heterozygous yeast deletion mutants to reveal genes encoding proteins with high flux control coefficients.

For the bottom-up approach, the initial problem is one of systems identification. While a lot of time is currently spent debating the question “What is Systems Biology?”, why (in an organism where we know so much about its biochemistry, physiology, and cell biology as S. cerevisiae) should it be a problem to identify the biological sub-systems that must be fully characterised and built into a comprehensive model of the eukaryotic cell? This problem arises because we have previously studied these biological systems in isolation and in a rigorously reductionist fashion. Now, we must study them as parts of an integrated whole. The problem is that our current view of, say, a metabolic or signal transduction pathway is often two-dimensional (rather than four-dimensional) and is frequently poorly integrated, if at all, with other cellular pathways. Thus our view of the network of metabolic pathways may not be the same as the yeast’s. In order to gain a “yeast’s eye view”, we have coupled flux balance analysis with both metabolomics and genetics. Although the initial aim of these approaches is the identification of the ‘natural’ metabolic systems of yeast, the principles involved should be more widely applicable to the problem of biological systems identification.

October 10, 2 pm

“System level analysis and engineering of industrial bacteria”
Sang Yup Lee, KAIST

- Choose two genes from microarray late stage → rise metabolite production
- Leptin production – Serine-rich production increase interlukin
- Enhanced production of recombination protein (patent)
- Silver Cell research at MBEL and Bic
- **http://webcell.org**
- MetaFluxNet v1.8
- Succinic acid productin increased by 4o times (US$ 550 million market)

References

[1] Appl. Environ. Microbial (2003), 69, 5772
[2] Trends in Biotechnoloy (2005), 23, 349
[3] Curr. Opin. Biotech (06), 17, 488

"Metabolome Analysis and Synthetic Biology"
Masaru Tomita (Keio University)

Keio University was founded by Yukichi Fukuzawa (appeared on the ￥10000 dollar notes)

- Metabolome analysis of AAP hepatoxicity in mouse liver
- Multi-omics → synethic biology
- Metabolic – CE/TOF-MS
- Fluxome – GC/MS – NMR, GC/TOF
- Proteome – shotgun, 2D gel
- Transcriptome – RT-PCR
- Merge two genomes – Bacillus
- Artificial operons – order of genes is important, Itaya et al.
- Metabloome factory – Tsuruoka
- *In vitro* enzyme rate constant (usually work at the maximum rate) is not equal *in vivo*

References

[1] PNAS (2005) 102, 15971

"A Systems Biology Approach to Identify and Therapeutically Exploit the Weakness of the Robust Tumour Metabolism"
Marta Cascante (University of Barcelona)

- Genomoics, transcriptomics, proteomics, interactomics, fluxics, metabolics
- Metabolic adaptation support cell function
- Design metabolites intercention in drug development need metabolic flux map
- Isotopomer analysis plus kinetic model
- Metabolic changes associates to cell transfer induced by oncogenes
- Disrupt metabolic to test the robustness of cancer cell

References

[1] Trends in Biotech. (2005) 16, 350.

<u>Oct. 11 9:15am</u>
Jack Donald Keene (Duke University Medical Center)
"Coordination of Gene Expression by RNA Operons"

- DNA operons and regulons
- Bacteria lack nucleus → efficient in transcription and transcription
- RBPS – S. cerevisiae 200 TFs, 560+ RBP, H. sapiens 1500+TFs, 3500+ RBP
- Robust, resilent, rewireable
- Intra-pathway PPI, inter-pathway PPI
- Coordination of PT gene expression
- Polycistronic operons

References

[1] Keene and Tenenbaum Mol. Cell (2002)
[2] Sci. (2005), Sep. 2.

Kenneth Alan Loparo (Case Western Research University)
"Applications of Complex Systems Biology to the Study of Neural Systems "

- Plasticity and activity dependent development
- Neural plasticity is a dynamic process by which the brain develop
- Affect by nature (GP), nurture (environment stimulus), niche (e.g. development)
- Brain development → first five years are most critical
- Synapses first three years, connect neurons
- Gene → phenotype ← environment (EGG, heart rate, blood pressure, temperature)
- By NMR imaging → connective
- Measure of brain connectivity and complexity using EEG
- Brain as a dynamics system
- Attractor geometry → periodic, Quasi-period, chaotic
- Correlation intergral, dimension of the integral has a power law
- Unfold the attractor
- Embedding the attractor
  - mapping that preserve the information
- plasticity an activity dependent dynamics process in decoupling organism

<u>Oct. 11　11:00 – 12:30 am</u>

Control and System Theory for Systems Biology - Oct. 11th, 11:00-12:30

"Robustness Analysis of Biological Networks Using Sensitivity Measures"
Francis J. Doyle (University of California, Santa Barbara)

- The wisdom of the body 1932
- Nebert Wiener 1948
- Mr. Black at IBM, develop negative FB loop device
- Chemotaxis, HSP, MAPK, signal transduction
- Attributes - intrinsic – unmodeled effect – transcription, transcription signal transduction, extrinsic – disturbance
- Analysis – sensitivity analysis – determine the change δ induced by initial condition on the prediction model (ODE, PDE)
- Linear analysis – positive or negative FBL, redundant loop, time delay, gain modulator, hierarchical, multi-scale system analysis (length and time)
- Need a metric for robustness performance
- Phase as a metric – do phase sensitivity analysis (Pittendrigh and Gaan 1976)
- Need to consider stochastic performance
- Isolated cell are terrible clock !

References

[1] Begheri, Stelling and Doyle, Biol. Rynthm Res.
[2] Mirsky, Stelling and Doyle (2006)
[3] Herzog (2004)

"Feedback Control Regulation of Cell Division"
Pablo A. Iglesias (Johns Hopkins University)

- Chemotaxis –a temporal sensing strategy (smell, high concentration, move)
- 6 to 10 flagellar/cell → rotatory motor
- Uses a biased random walk, clockwise → flagellar apart then stay randomly, anti-clockwise → flageller come close and travel in a straight line
- Barkai, Leibier paper → property of perfect adaptation is robust, initial concentration increased by 50 times the system still stay robust
- To determine the direction of move, need to known the rate change of concentration with time
- Kalman filter theory

"The architecture of cellular regulation"
John Doyle (California Institute of Technology)

- Rule of interaction is not module
- Thermodynamics, communication, control and computation
- Few polymerase which can eat ~20 same things (no variety)
- ~10000 proteins give rise to huge variety
- Robust and yet fragile
- Flexible metabolism obesity and diabetic
- Regenerate　　　cancer
- Advance technology castrographic
- Virus NFkB, toxity
- Bowtie architecture – long-time and short-time scale
- Computing protocols for evolution
- Evolving of architecture (like internet)
- Circadian clock in a bacteria
- How to compare architecture

Oct. 11th, 14:00-16:30

Signal Transduction -

Chairs: Philippe Bastiaens (EMBL Heidelberg), Boris Kholodenko (Thomas Jefferson Univ.), Shinya Kuroda (Univ. of Tokyo)

"Emerging Principles of Living Systems "
Hans V. Westerhoff (The University of Manchester)

- Emerging principle is a theory for tradeoff between robustness
- Robustness is conserved ?
- Definition of robustness ?
- Robustness and fragility
- Robustness is a function
- Definition – frequency domain (phase)
- Homogeneous robustness – average robustness is lower
- Heterogeneous robustness – average robustness is lower
- Another question is to which is this perturbation that signal transduction pathways change ?
- Normal fragility to cancer to robustness

William S. Hlavacek (Los Alamos National Laboratory)
"Rules for Modeling Signal-Transduction Systems"

- Problem of combinatorial complexity
- Rule-based model
- EGFR multiplicity of site and binding partners to combinatorial complexity
- For example: 9 sites is equal to 2^9 = 512 sites
- Each site has more than one binding partner, that is 3^9 states
- Protein inhibits dimmer breaking, no protein monomer
- Develop a tool – BioNetGen2
- Five proteins in EGFR (edge bonding is either intra or inter bonding)
- Epitope – trivalent ligand with a bivalent cell surface receptor
- 23 reactions, 21 metabolites can give rise to 622 isotopomer fraction

References

[1] Faeder (2005) Proc. ACM

[2] Blinov (2005) Proc, BopCONCUR

[3] Blinov (2006) Biosystems

Philippe Bastiaens (EMBL Heidelberg)
"Reaction cycles in the spatial and temporal organization of cell signaling"

- RAN GTP gradient in the self-organization of mitotic spindle
- Spatial heterogeneity of tryosin receptor

$$E + S = ES \rightarrow E + P$$

- Advantages are specificity, localization, intrinsic kinetic and parameters
- Measure ES by FRET in vitro
- Imaging of ES in cell
- Observed the complex of ES for a long time by depolarization the complex
- Steady state of ES in cells → $k_1(1-\alpha)S - (k_{-1} + k_2)\alpha = 0$
- k_2 is the same as k_{cat}
- $K_m = (k_{-1} + k_2)/ k_{-1} = (1 - \alpha)S/ \alpha$
- Signal permission region and signal terminating region

三、考察參觀活動(無是項活動者省略)

Null

四、建議

In summary, the workshop was taken place at the Pacifico Yokohama building at Yokohama, and had a lot of discussions. The level and quality of the talks are very good. The talk gave by Marc Vidal is extremely good. I strongly recommend one should understand his work as thoroughly as possible if anyone want to do system biology research. His work is a ground-breaking work.

五、攜回資料名稱及內容

資料名稱:

(1) ICSB-2006 programme

(2) ICSB-2006 abstract CD

六、其他

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

96 年 1 月 29 日

附件三

報告人姓名	吳家樂	服務機構及職稱	亞洲大學 生物科技與生物資訊系 副教授
會議 時間 地點	14-17 Jan. 2007 Hong Kong	本會核定補助文號	NSC 95-2745-E-468-008-URD
會議名稱	(中文) 2007 年亞太生物資訊研討會 (英文) The Fifth Asia Pacific Bioinformatics Conference 2007		
發表論文題目	(中文) 蛋白質區域與區域作用於生物路徑之應用 (英文) Application of domain-domain interaction in pathway study (中文)人類 miRNA 與起動子區及 CpG 區關聯之預測 (英文) Predicting Putative Human Mirna Precursor Candidates Associated with Promoter Regions and CpG Islands		

報告內容應包括下列各項：

七、參加會議經過

Jan. 14

I attended three tutorials, which is listed in the following;

- Tutorial 1: Guilt by Association: A Tutorial on Protein Function Inference
  Prof. Limsoon Wong
- Tutorial2: Clinical Proteomics and Biomaker Discovery - Usage and Abusage of Bioinformatics Tools
  Prof. Zhen Zhang
- Tutorial 3: Introduction to Phylogenetic Networks
  Prof. Daniel H. Huson

Jan. 15 attended session 1 – communities, motifs, and session 2 – biclustering and the micro-array talks

Jan. 16 attended session 4 –structure prediction and comparison, and session 5 – mapping and disease talks
*Poster presentation at 15:00 p.m.:

Jan. 17 attended session 7 –biological network, and session 8 – MS, protein-protein interaction talks

八、與會心得

The Asia-Pacific Bioinformatics Conference 2007 is an annual forum for exploring research, development and novel applications of Bioinformatics which is held at the University of Hong Kong, from Jan. 14 to 17.

The scientific program of APBC 2007 included 3 keynote talks, 3 tutorials, 35 oral presentations, 112 poster presentations and a HP industrial sessions as well.

The symposium has received 104 papers and each submitted paper was reviewed. All accepted papers had at least 2 positive recommendations. The program committee accepts approximately 33% of papers, that is a total of 35 papers. A variety of papers was presented at this conference and the topics include protein structures study, motif search, micro-array analysis, proteomics, pathways, networks and evolution study.

I had two poster presentation on Jan. 16, 15:00 p.m. Title of my two posters presentation are "Applications of Domain-Domain Interaction in Pathways Study ", and "Predicting Putative Human Mirna Precursor Candidates Associated with Promoter Regions and CpG Islands ". On the other hand, I had attended most of the talks during the four days conferences.

In my personal opinion, bioinformatics researches are growing very rapidly, and it is moving into the areas such as data integration, gene micro-array analysis, proteomics and system biology. There are several good tutorials, talks and posters presented in this conference. I will highlight their main interesting results in below.

Jan. 15, 2007

Keynote: Exploring Genomes of Distantly Related Mammals

Prof. Jennifer A. Marshall Graves
Professor
ARC Center for Kangaroo Genomics, Research school of Biological Sciences
Australian National University

- Physical mapping – FISH, BAC, orthologs
- Linkage mapping – markers
- Evolution – MHC locus, framework gene class (I, II, III)
- Marsupial – anti-body, milk-commercial products

- Centerome (few recombination)

session 1 – communities, motifs talks

Metagenome Analysis using Megan
by Daniel Huson, Alexander Auch, Ji Qi, Stephan Schuster
Subtle Motif Discovery for Detection of DNA Regulatory Sites
by Matteo Comin, Laxmi Parida

- Metagenomics – study of the DNA of uncultured organism
- Sargasso sea – 1.2 million new genes

Algorithmic Approaches to Selecting Control Clonesin DNA Array Hybridization Experiments
by Qi Fu, Elizabeth Bent, James Borneman, Marek Chrobak, Neal Young

- rRNA control probe
- BCP – back covering problem is NP complete

Subtle Motif Discovery for Detection of DNA Regulatory Sites
by Matteo Comin, Laxmi Parida

- Make use of MSA of TFBS
- Take into account of noise, such as indel and mutation
- Take a TFBS with length 15, 4 mutations for each 600 bp
- Known approaches include: exact enumeration schema, inexact (sub-motif enumeration), combinatorial algorithm (map to a clique problem), statistical learning problem (gibbs sampling problem), indel consensus problem
- Modular approach, step 1: identify potential signal which allow variable length (PROSITE), step 2: local search of gap length is variable, that is alignment of potential signal
- Problem of extensible motif can be solved by assigning a probability

An Effiective Promoter Detection Method using the Adaboost Algorithm
by Xudong Xie, Shuanhu Wu, Kin-Man Lam, Hong Yan

- The authors start with using 5-mer (that is 1024 possibilities)
- Apply Bayer rule
- Rank the occurrence of 5-mer

1 CGGCC
2 GCGCG
3 GCGGC
4 ..
5 ..
…

- 65% of binding sites are CG rich regions
- Several features can be employed to do the search
- Number of 5-mers
- Position of CpG islands
- DNA sequences
- Test of capability of exact TSS localization
- Define (GC)p = %C + %G > 50%
- Experimental to observe ration > 0.6 for a region longer than 200 bp

session 2 – biclustering and the micro-array talks

A New Strategy of Geometrical Biclustering for Microarray Data Analysis
by Hongya Zhao, Alan W. C. Liew, Hong Yan

- How to classify a subset of genes and a subset of conditions
- NP hard problem
- First do a bicluster (Br) in column pair space, then merge Br into maximal cluster

Using Formal Concept Analysis for Microarray Data Comparison
by Vicky Choi, Yang Huang, Vy Lam, Dustin Potter, Reinhard Laubenbacher, Karen Duca

- FCA is first proposed by Rudolf Wille, see FCA: mathematical concepts
- Consider objects (girl, women, boy) with certain attributes (different names but same things)
- Concept lattice, bipartite clique

	1	2	3	4
a	0	0	1	1
b	1	0	0	1
c	1	0	1	0

d.....

- bipartite clique {(a), {1,3}}
- maximal bipartite clique ({a,c},{ !,3})
- partial order – Galois lattice)
- comparing lattices
- gene expression + PROSITE motifs
- 8 discrete gene expression values and 21 PROSITE motifs
- Common sublattices
- Define global differential expression
- Lattice decomposition and sublattices – more biological related attributes

An Efficient Biclustering Algorithm for Finding Genes with Similar Patterns in Time-series Expression Data
by Sara Madeira, Arlindo Oliveira

- 3 biocluster ➔ 3 biological processes
- S = (D N U) = (down no-change up)
- Generalize suffix tree
- Maximal = CCC-bicluster, (left, right) maximal
- Gene = (g1, g2 g3 …), condition = (D N U …..), motif = (D1, N2, U3, U4, N5)

Selecting Genes with Dissimilar Discrimination Strength for Sample Class Prediction
by Zhipeng Cai, Randy Goebel, Mohammad Salavatipour, Yi Shi, Lizhe Xu, Guohui Lin

- Similar gene expression level does not give useful information
- Gene selection methods carefully define a function to score the differential levels of gene expression inder a varity of conditions, in order to identify top-ranked genes
- Such method suffer because some genes have very similar expression patterns so using them all in classification is largely redundant
- These genes can prevent the consideration of other individually-less but collectively-more differentially expressed genes
- The authors proposed to cluster genes in terms of their class discrimination strength and to limit the number of selected genes per cluster
- The authors showed by experiments on two cancer microarray datasets that their methods identify gene subsets which collectively have significantly higher classification accuracies

Attended the poster session.

Jan. 16, 2007

Keynote: Protein Identification via Spectral Networks Analysis

Prof. Pavel Pevzner
Ronald R. Taylor Professor of Computer Science
Department of Computer Science & Engineering
University of California, San Diego

- PTM is important, more than 600 PTM are known
- Positive identification via spectral network analysis
- Human eyes lens – cataracts
- PTM of aged protein accumulate
- GPFN cot fragments (kDa)
- Spectrum of mass + noise
- Cut at peptide bond, H2O, NH3
- Identify position and intensity
- A modified petide at most k mutation/modification apart from the database
- Compare 1 million spectra with database
- Search the MS/MS without look at 99.9% of the database – find similarity
- Search the MS/MS without compare the spectra database
- Denote PTM (increase the mass in MS/MS), use zero to denote such peak
- Edit distance/alignment problem (prefix/suffix)
- Spectral alignment = east-west TSP
- De novo without error
- Identify 6 new PTM
- PTM ➔ polymorphism (A-Bruijn graph)

session 4 –structure prediction and comparison talks

Protein Structure-Structure Alignment with Discrete Frechet Distance
by Minghui Jiang, Ying Xu, Binhai Zhu

- Treat protein sequence as a polygonal chain (use of Frechet Distance, FD)
- Hausdorff distance is useful for matchibg two point sets
- The discrete Frechet distance closely approximates the (continuous) Frechet distance and is a natural measure for the geometric similarity of the folded 3D structures of proteins
- New algorithms are proposed for matching two polygonal chains in 2D to minimize their discrete FD under translation and rotation and an effective heuristic for matching two polygonal chains in 3D structure-structure alignment
- FD can be computed by dynamic programming |Bj| > |Ai| = 1, where A and B are two sets of points

Deriving Protein Structure Topology from the Helix Skeletion in Low Resolution Density Map using Rosetta
by Yonggang Lu, Jing He, Charlie Strauss

- Electron cryo-microarray is an experimental technique to determine the 3D structure for large protein complexes
- Able to generate protein density maps at 6 to 9 A
- Secondary structures such as alpha-helix and beta-sheet can be visualized from these maps, but there is no mature approach to deduce their tertiary topology, the linear order of the secondary structures (SS) on the sequences
- Given N SS, the number of possible orders is $N!(2^N)$
- The authors develop a method to predict the topology of the SS using *ab initio* structure prediction

- The rosetta structure prediction algorithm was used to make sequence based structure predications for the protein
- Then screened the models produced by Rosetta for agreement with the helix skeleton derived from the density map
- For most of those true-positive assignments, the alignment was accurate to within 2 amino-acids in the sequences

Fitting Protein Chains to Cubic Lattice is NP-Complete
by Jan Manuch, Daya Gaur

- It is known that folding a protein chain into the cubic lattice is an NP-complete problem. Given a 3D fold of a protein chain (the alpha carbon coordinates), the authors want to find the closest lattice approximation fo this fold. This problem has been studied under names such as “lattice approximation of a protein chain”, “the protein chain fitting problem” and “building protein lattice models”. The authors show that this problem is NP-complete for the cubic lattice with side 3.8A and the coordinate root mean-square deviation.

Inferring a Chemical Structure from a Feature Vector Based on Frequency of Labeled Paths and Small Fragments
by Tatsuya Akutsu, Daiji Fukagawa

- This paper proposes algorithms for inferring a chemical structure form a feature vector based on frequency of labeled paths and small fragments, where this inference problem has a potential application to druh design.
- Chemical structures are modeled as trees or tree-like structures.
- The inference problems for these kinds of structures can be solved in polynomial time using dynamic programming-based algorithms. A branch-bound type algorithm is also proposed.
- The result suggests that the algorithm can solve the inference problem in a few or few-tens of seconds for moderate size chemical compounds.

session 5 – mapping and disease talks

Exact and Heuristic Approaches for Identifying Disease-Associated SNP Motifs
by Gaofeng Huang, Peter Jeavons, and Dominic Kwiatkowski

- Some combinations of SNPs in the human genome are known to increase the risk of certain complex genetic diseases. The authors formulates the problem of identifying such disease-associated SNP motifs as a combinatorial optimization problem and shows it to be NP-hard. Computational results are given to demonstrate that these approaches are sufficiently effective to support ongoing biological research.

Genotype-Based Case-Control Analysis, Violation of Hardy-Weinberg Equilibrium, and Phase Diagrams
by Young Ju Suh, Wentian Li

- The author study in detail a particular statistical method in genetic case-control analysis, labeled genotype-based association, in which the two test results form assuming dominant and recessive model are combined in one optimal output.

Attended the poster session.

Jan. 18, 2007

Keynote: Bugs, Guts and Fat - a Systems Approach to the Metabolic 'Axis of Evil'

Prof. Joe Nadeau
Chair/Professor
Department of Genetics
School of Medicine
Case Western Reserve University

session 7 –biological network talks

Infering Gene Regulatory Networks by Machine Learning Methods
by Jochen Supper, Holger Fröhlich, Christian Spieth, Andreas Dräger, Andreas Zell

- The authors critically evaluate the application of multiple linear regression, SVMs, decision trees and Bayesian networks to reconstruct the budding yeast cell cycle network. The performance of these methods is assessed by comparing the topology of the reconstructed models to a validation network. This validation network is defined as a priori and each interaction is specified by at least one publication. The author also investigate the quality fo the network reconstruction if a varying amount of gene regulatory dependencies is provided a priori.

A Novel Clustering Method for Analysis of Biological Networks Using Maximal Components of Graphs
by Morihiro Hayashida, Tatsuya Akutsu, Hiroshi Nagamochi

- The authors proposed a novel clustering method for analyzing biological networks, in this method, each biological network is treated as an undirected graph and edges are weighted based on similarities of nodes. Then, maximal components, which are defined based on edge connectivity, are computed and the nodes are partitioned into clusters by selecting disjoint maximal components. The proposed method was applied t clustering of protein sequences and was compared with conventional clustering methods. The obtained clusters were evaluated using p-values for GO terms, the average p-values for the proposed method were better than those for other methods.

Gene Regulatory Network Inference via Regression Based Topological Refinement
by Jochen Supper, Holger Fröhlich, Andreas Zell

- Starting from a priori specified network topologies, the authors identify those parts of the network which are relevant for the gene experiment data. For this propose, the authors employed linear ridge regression to predict the expression level of a given gene from its relevant regulators with high reliability. Calculated statistical significances of the resulting network topologies reveal that slight modifications of the pruned regulatory network enable an additional substantial improvement.

Algorithm Engineering for Color-Coding to Facilitate Signaling Pathway Detection
by Falk Hüffner, Sebastian Wernicke, Thomas Zichner

- To identify linear signaling pathways, Scott et al . recently proposed to extract paths with high interaction probabilities from protein interaction networks. They used an algorithmic technique known as color-coding to solve this NP-hard problem, their implementation is capbable of finding biologically meaningful pathways of length up to 10 proteins within hours. In this presentation, the authors give various novel algorithmic, improvements for color-coding. Experiments on the interaction networks of yeast and fruit fly as well as a testbed of structurally comparable random networks demonstrate a speedup of the algorithm by orders of magnitude.

session 8 – MS, protein-protein interaction talks

De Novo Peptide Sequencing for Mass Spectra Based on Multi-Charge Strong Tags
by Kang Ning, Ket Fah Chong, Hon Wai Leong

- The authors presents an improved algorithm for de novo sequencing of multi-charge mass spectra. A simple de novo algorithm, called GBST (greedy algorithm with best strong tag) was proposed and was shown to produce good results for spectra with charge larger than two.

Complexities and Algorithms for Glycan Structure Sequencing using Tandem Mass Spectrometry
by Baozhen Shan, Bin Ma, Kaizhong Zhang, Gilles Lajoie

- The authors showed that glycan de novo sequencing is NP-hard. The authors provide a heuristic algorithm and develop a software program to solve the problem in practical case. Experiments on real MS/MS data of glycopeptides demonstrate that the authors' heuristic algorithm gives satisfactory results on practical data.

Semi-supervised Pattern Learning for Extracting Relations from Bioscience Texts
by Shilin Ding, Minlie Huang, Xiaoyan Zhu

- The authors proposed a semi-supervised model to combine both unlabeled and labeled data for the pattern learning procedure. First a large amount of unlabeled data is used to generate a raw pattern set. Then it is refined in the evaluating phase by incorporating the domain knowledge provided by a relatively small labeled data. It is showed that labeled data when used in conjunction with the unlabeled data can considerably improve the learning accuracy.

Flow Model of the Protein-protein Interaction Network for Finding Credible Interactions
by Masanori Arita, Kiyoshi Asai, Kinya Okada

- The authors proposed the relative reliability score for protein-protein interaction (PPI) as an intrinsic characteristic of global topology in the PPI network. The score is calculated as the dominant eigenvector of an adjacency matrix and represents the steady state of the network flow. By using this reliability score as cut-off threshold from noisy Y2H PPI data, the credible interactions were extracted with better or comparable performance of previously proposed methods which were also based on the network topology.

九、考察參觀活動(無是項活動者省略)

無

十、建議

In summary, the symposium had a lot of discussions. The level and quality of the talks are very good. The organizer had done a very good job in organizing the conference.

十一、 攜回資料名稱及內容

資料名稱:

(1) Proceedings of the Fifth Asia-Pacific Bioinformatics Conference, and

(2) Poster abstract - The Fifth Asia-Pacific Bioinformatics Conference.

十二、 其他

無

附件三

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

96 年 7 月 31 日

報告人姓名	吳家樂	服務機構及職稱	亞洲大學 生物科技與生物資訊系 副教授
時間 會議 地點	21-25 July 2007 ISMB/ECCB 2007 Vienna, Austria	本會核定 補助文號	NSC 95-2745-E-468-008-URD
會議 名稱	(中文)第十五屆智慧系統分子生物學 2007 與第六屆歐洲計算生物學國際研討會 (英文) The 15th International Conference on Intelligent Systems for molecular biology, 6th Annual European conference on computational biology		
發表 論文 題目	(中文) 鄰近於 CpG 島之 microRNA 基因 (英文) CpG-islands-proximal MicroRNA Genes		

報告內容應包括下列各項：

十三、　參加會議經過

July 22

- Attended the Keynote session
- Attended the morning sessions
- Attended the afternoon sessions

July 23

- Attended the Keynote session
- Attended the morning sessions
- Attended the afternoon sessions

July 24

- Attended the morning sessions
- Attended the afternoon sessions

July 25

- Attended the Keynote session
- Attended the morning sessions
- Attended the afternoon sessions

十四、　與會心得

This year the ISMB/ECCB07 conference was held at Vienna, Austria The conference comprises a broad area of topics in the area of computational biology.

Topics include,

- Protein-protein interaction (PPI) systems biology and pathways reconstruction
- microRNA, nc-RNA
- MAPK Signal transduction pathways
- Integration of microarray data, genome sequences, and PPI data
- PPI and evolution
- Host-pathogen PPI prediction
- Systems Neuroscience

I had a poster presentation from July 22 till 25. The title of my poster presentation was CpG-island proximal microRNA genes

There were many interesting and good talks presented in this conference. I highlighted their main interesting results in below.

July 22 8:30 am
“Dissecting transcriptional network structure and function”
Erin K. O’Shea

- Under transcription factors (TFs) regulation mechanisms
- TFs – TFBS have different sensitivity, cooperative levels

```
                     TF1            TF2
                 |   |   |       |   |   |
   Signal_1   +   +   0       +   +   -
   Signal_2   -   +   0        -   +   -
Signal_1_2    +   +   0       +   +   -
```

➔ build transcription network

- Structure with a signal and trace the influence through TFs
- Mutant cycle approach – (i) measure the influence of each TF with significant and (ii) identify a quantity
- Yeast – salt stress and TM stress, involve the HOG1 TF
- In cytoplasm – HOG1 translocate into nucleus where it affects a lot of other TFs, such as msn2/4, sko1, hot1,msn1, smp1 → up/down about 300 genes
- The question they want to solve is to measure the cooperative effect of signal one and two, which is not simply the sum of both signal, that is a non-linear effect
- The effects (stress response) can be described by the Boolean logic language – OR, AND gate

July 22 9:30 am
“Inferelator: learning predictive dynamic regulatory networks from heterogeneous data”
R. Bonneau

System architecture – cMonkey, do a bicluster, find bicluster motif, combine with Inferelator to predict the gene regulation network (GRN)
Co-regulator information is better than co-expression information
Associated regulon
cMonkey – make use EM algorithm, (i) do a single bicluster iteration, (ii) solve the differential equation for steady state (use linear model or logistic model)

July 22 10:00 am
"Redefining nodes and edges: relating 3D structures to protein networks provides evolutionary insights"
P. Kim, Yale University

- PIN – determined by TAP-tagging, Y2H experiments
- Hubs – data hubs and party hubs
- Data hubs – interact at different space and time
- Party hubs – co-express at the same space and time
- Hubs are essential and slow evolve
- Define the single-interface network (SI), multi-interface network (MI)
- Do a PDB homology map of the PIN
- Degree centrality – different subcellular localization (SL)

July 22 10:50 am
"Identification of functional modules from conserved ancestral proten-protein interactions"
J. Dutkowski

Comparative genomic – across species, there are conserved subnetworks
Cluster – protein complexes
Conserved interaction modules – evolve together
Path length – common signal network
Clustering of protein with high e-value of BLAST
Model – duplication (local), speciation (global)
Define the probability of interaction, and probability of new species, then construct the Bayersian network

July 22 11:20am
"Functional annotation of regulatory pathways"
M. Koyuturk, koyuturk@gmail.com

- Study GRN in A. thaliana, flowering time
- Synthetic gene array network (SGA), Tong et al . Sci. 2004
- From gene space to function space
- Define topological parameters in function space and build the functional network
- Reduce complexity – short circuit TFs, DNA binding nodes, and get common functional attributes
- NARADA – **http://www.cs.purdue.edu/homes/jpandey/narada**
- One can also consturcut GRN from the Regulon DB
- The idea is from function-function network, do a short circuit, get a new view of the GRN

July 22 11:50 am
“supervised reconstruction of biological networks with local models”
K. Bleakley

- Unsupervised method, model based graph, do a BLAST search
- Supervised method – SVM
- Assign a 23 bit vector to a protein, such as similarity features
- Build a network – functional related (+1), fixed one node and do the learning process
- Put a boundary between functional related and not functional related network
- Advantages of local model – fast, focus on a subnetwork, easy extend to direct graph
- Disadvantages of local model – two new protein no simple way to predict there is an edge
- False discovery rate (FDR) – among the positive which one is negative (false)
- The model has a very good FDR power, for example among 20000 edges only 1000 edges is false
- Strong positive for FP is good candidate for missing

July 22 1:30 pm - Keynote
"understanding interactions by data integration"
S. Brunak (Technical University of Denmark)

- Beyond one gene cell cycle, cell cycle protein have very different sequences but interact at the same phase
- Benchmark data before integration
- Temporal interaction network are just in time, adding cell cycle metabolome data
- Protein degradation
- What do species do in between regulation, transcription regulation (TR) is poorly conserved

- Protein complex activate at the right time
- TR is conserved at the level of protein complex rather than conserved at single protein
- Disease interactomes – ranking protein complexes, identify new disease genes
- Phenome – interactome network

July 22 2:30 pm

"An ensemble framework for clustering PPIN"

S. Asur

- Non-uniform degree distribution is hard to apply clustering
- Ensemble clustering – based on different topology to classify
- Different criteria clustering – single cluster
- Topology parameter – clustering coefficient
- Shortest path betweenees is global topological properties
- Consensus clustering – agglomerative, soft hierarchical clustering allow multiple classes
- Topological modularity, domain-based measure
- Validate the measure
- PCA based consensus method is better than other type of classifying methods

July 22 3:50 pm

"EMBOSS" by EBI

- A EBI open software suite

July 22 4:20 pm

"comparative genomics of translation regulation in yeast "

P. Yitzhak

- Study tRNA abundance verses tRNA gene copy
- Build a translation efficiency matrix
- Look at the ribosomal protein
- Study the yeast frozen tRNA pool
- Study fungi – such yeast, S. pombe
- Relative level of translation/relative level of transcription ~ noise residue

July 22 4:50 pm

"a graph-based approach to systematically reconstruct human transcriptional regulatory modules"

X. Yan

- Coexpression network – a multigraph problem simplified to a single graph with weight on the edges
- Summary graph – noise edge problem, due with the bionomial distribution
- Graph = random graph + real graph
- NeMo – a graph-based approach

July 22 5:20 pm

"systematic discovery of functional modules and context-specific functional annotatin of human genome"

Y. Huang

- EBI ArrayExpress ~ 55000 experiments, GEO ~ 137231 experiments
- Multiple samples – reduce noise (gene expression levels)
- Problem – different file formats
- From microarray data to coexpression network to recurrent pattern to functional annoation
- Problem is the 65 microarray data set has more than 1 million rows

- Simplify by define S = p + p*p + p*p*p + …..
- If less than 3, it is a local area

<u>July 22 5:50 pm</u>
"understanding and expliting the evolution of the sequences that control gene expression"
M. Eisen (UC, Berkeley)

- Binding site clusters conserved is more useful
- TFBS – mutation is easy to erase their function
- Fly regulatory sequences are amazing pastic
- The TFBS they contain can be completely rearrangement but the function still remain
- Consider enhancer elements, miRNA genes
- Conservation is not equal to funciton

<u>July 23 8:45 am</u>
Interaction networks probed by mass spectromeety
Anne-Claude Gavin, EMBL

- Study PPI in Hs. Syndromes
- Protein could has many functions – genetic pleiotropy
- Near every process is carried by more than ten protein subunits – protein complex also conserved, such as ARP2/3 complex
- Protein complexes are dynamic (TNFa), drug induced complexes
- Predict protein complex – cooperative effects (not just binary information), alleosteric effect
- Study protein complexes in Sc. (yeast)
- Qt – SNARE complex
- Quantity protein social affinity, two models; (i) spoke model, (ii) matrix model (Core component with attachment)
- Modules – groups of protein present in more than one complex
- Discrete organization – architecture of protein complex
- A molecular framework of phenotype data
- Mutation of protein complex gives same phenotype
- Conclusion: (1) more than 80% proteins are belong to protein complexes, (2) a protein complex has more than 40% proteins belong to a protein complex, and (3) social affine index is a useful measure

<u>July 23 9:30 am</u>
Genetic networks: inferring pathways by computational perturbation
F. Roth

- Synstematic genetic inference
- Study DNA damage, identify mutant sensitivity to MMS, measure the cell thickness
- Define fitness(W)=growth rate in mutant/growth rate in WT=doubling time in mutant/doubling time in WT
- They pick the product model, define e = Wxy – WxWy
- Non-interacting, e = 0
- Synergenetic e < 0
- Alleviating e > 0
- Classify alleviating interaction by drug response, S = response with drug/response without drug
- Final link to GO terms
- Genetic interaction prediction from relationships

- Reconstruction of DNA repair pathway
- Most functional links are alleviating – pathway order
- Co-equal interaction to cohesive pair
- Alleviating interaction implies pathway order

July 23 10:00 am

Domain-domain interactions (DDI) are evolutionary conserved
Z. Itzhaki

- Reference: BMC genome biology, 7, R125
- Protein PSM has a domain PDZ, IL16 has two domains, PDZ, PDZ
- Unstructured region (Transient interaction) and structure motif (stable interaction)
- Databases: 3D1D, iPfam, an overlap of 2983 DDI
- Statistical analysis of DDI
- Create 1000 random PPI network and count the PPI due to DDI, then set a p-value
- BioGrid + DIP + InterAct
- Study the following speces: Ec, Sc, Ce, Dm and Hs
- Why some DDI are more conserve, others are less
- DDI are more conserved than PPI
- Estimate the number of DDI non-conserved due to incomplete PPI

July 23 1:30 pm

Dissecting transcriptional network structure and function
Eran Segal

- Develop a mechanistic model of transcription control (TC)
- Cluster of TFBS (weak or strong), cooperative, quenching
- Fly model (maternal genes, gap genes, pan-rule genes)
- How CRM compute expression pattern ? Input – different combinations – different expression patterns
- Integrate factor concentration and binding energy
- Model binding competition through steric hinderance
- Allow contribution from both weak and strong sites
- Two parameters of the models: absolute concentration parameter, expression concentration parameter
- Predict expression profile, measured profile (fitting parameter for all CRM)
- Only account for 80% of the cases
- Given the sequence to predict the expression profile
- A model of design principle of segmentation
- Are modules densely populates with factor ? both strong and weak sites contribute
- TFBS same TF tend to cluster
- Nucleosome position is important, di-nucleotide AA/TT/TA for DNA binding
- Nucleosome signal is highly conserved across species

July 23 2:30 pm

The relationship among sequence diversity, coevolution, and specificity in protein interactions
S. Lovell

Reference: PNAS (2007) 104, p. 7999
The plot of shared interaction against sequence identities show that there is no correlation al tall
Pairwise with one member show accelerated evolution
PDB to SCOP, co-evolution and co-adaptation

Do MSA – construct the mirror tree
Study protesome, co-evolution, correlated evolution, site-specific co-evolution

July 23 3:50 pm
Large scale mapping of human protein-protein interactions by mass spectrometry
T. Topaloglou

- 6400+ PPI among 22000 proteins
- Reference System biology 3, article 89
- Data available InterAct (EBI-1050370)
- Built cDNA – cell IP – LC-MS/MS – data generation – PPI network
- Author used spoke model, so less FP (a flow chart process to filter FP)
- Interaction confidence score, MS/MS PPI data verify
- Validation – other orthogonal source of PPI
- Biological process – bait-prey coincidence
- Node size = number of prey

July 23 5:45 pm
Why small RNA are highly coserved
John Mattick

- Devote to an RNA regulatory system to diversity, differentiation, and development
- ncRNA increase with organism complexity
- mRNA + intronic RNA implies functional
- how is regulation scale with function, $r = a n^b$, b = 2 is ideally
- it is found that r ~ n (1.96)
- reference Sci (2005), 307, 856
- ncRNA rapid evolution (see TIG (2006), 22, p1)
- lack of conservation not mean lack of function
- cis-antisense stand ncRNA
- function of ncRNA
- chromatin modification (DNA:DNA:RNA), (2) transcriptional regulation, (3) control of alternative splicing, (4) RNA modification and silencing

July 24 9:30 am
"RNA structure prediction, comparison and motif search "
J. Reeder

- Shape abstraction, ncRNA genes has no ORF information
- Recent advanced – 3D motif, chemical probing, Rfam as a gold standard
- RAVENNA – use shape abstraction from CM to HMM to speed up Rfam search
- ncRNA gene prediction, RNAz, EVOFOLD
- structure alignment as trees and points as string
- LOCOMOTIF – generate from thermodynamics matcher and present as grapher
- Shape abstraction (SA) – retain nesting and adjancy of stems
- Retain or disregard types of bulge or internal loops
- SA mathematics, SA function, shape representation structure, SA probability = Boltzmann probability of a shape
- RNA classification – predict alternative structure form consensus prediction
- Classification via shape dominance (probability larger than 0.5)

- Any SA function that is a tree homomorphism can be integrated into your DP program

July 24 10:00 am
"non-coding RNA"
P. Stadler

- Easy for ncRNA genes (not much indel), such as (r, t, tm, mi, SRP)RNA
- Hard for U7, SnRNA, yRNA, vault RNA
- Method – use Infernal then Locarna
- FRAGREP2
- RNAz – ncRNA gene finding tool, method – MSA + RNASS + MFE
- Predict Hs genome has about 36000 ncRNA
- ncRNA annotation problem – RNA family (high homology), RNA class (based on structure), RNA gene finding problem
- RNAstrand – on which strand is ncRNA located
- GC, GU substitution is different
- miRNA is rather conserved in mammalian
- putative novel RNA classes

July 24 9:30 am
"ncRNA motif detection "

- consider the ncRNA secondary structure, do a gap remove
- pairwise alignment, has a different meaning of local alignment
- MSA of ncRNA, no progress clustering allow consensus contain nested loop
- Tools : MARMA, locaRNA – clustering of RNAz ncRNA prediction
- MEMERIS – sequences motif with structure features

July 24 1:30 pm
“RNA folding with pseudoknot”

- Build a mechanical model of folding RNA structure
- Study the force extension curve of small RNA motif
- RNA knot - in co-transcriptional folding and RNA self-assemble
- Pseudoknoe versus knot, two ways to do it, (entangle helices, chirality)

July 24 2:00 pm
“influence flow: integrating pathway-specific RNAi data and protein interaction data”
R. Singh

- RNAi + PPI for MAPK signal transduction pathway prediction
- Define RNAi score
- Core cascade – high flow part, non-cascade – lower or less flow part
- Multi-commodity formulation is quite flexible
- Dealing with noise indicate FP (define a threshold to cutoff the RNAi signal)
- Epistasis, synthetic lethal experiments

July 25 8:45 am
“Genomic SELEX for the identification of novel non-coding RNAs independent of their expression level”
Renee Schroeder, Department of Biochemistry, University of Vienna

- Two ways to study RNA expression
- RNomics – isolate RNA – cDNA – size selection – c-tailing
- Microarray - isolate RNA – cDNA – binding
- Both use cDNA only see condition at one fixed time point
- Genomic SELEX – size selction – RNA pool in vitro – selection by looking RNA-protein complex
- Two examples of gene SELEX
- Isolation of Hfq binding RNA from Ecoli
- Hfq is essential for RNA-RNA binding
- Biding studies – primer do not affect RNA motif
- Do both (with primer – Hfq, without primer – Hfq)
- Anti-sense RNA regulate sense RNA
- D(anti-sense RNA) ~ D(sense RNA)
- D(anti-sense RNA) < D(sense RNA) - Control the transcriptional level
- Large cis-asRNA – there could be anti-sense pervasive translation

July 25 8:45 am
Knowledge management for modeling cell cycle control

- DIAMONDS
- CCO – cell cycle ontology web tool, **http://www.CellCycleOntology.org**
- Provide information, such as what (CyclinB), where (cytoplasm), when (interphase)

July 25 10:15 am
Computational prediction of host-pathoen PPI
M. D. Dyer

- Predict host-host (H-H), host-pathogen (H-P), pathogen-pathogen (P-P) PPI
- No gold standard database for validation
- Define a triplet state, that H-H-P, H-P-P
- Study triplet coexpression
- Functional enrichment

July 25 1:45 pm
Keynote
Computational biology: what is next?
Temple F. Smith

- Chomosky hierarch of protein domain
- Single domain – regular expression
- Concatenates domains – context free language
- Intercalate domains – context dependent language
- Interlace domains – regular enumerate language
- Pathway overlap control by compartmentalization
- Phenomme – filament growth, osmolarity response
- Species specific domain or block
- Ribosomal protein – ribosomal peptidyl transferase protein to answer the evolution problem

十五、　考察參觀活動(無是項活動者省略)

Null

十六、　建議

The conference was taken place at the Vienna international center at Vienna, and had a lot of discussions. The level and quality of the talks are outstanding. The talk gave by O'Shea is extremely enlightening. Her work is a ground-breaking work.

十七、　攜回資料名稱及內容

資料名稱:

(3) ISMB/ECCB07 programme
(4) ISMB/ECCB07 CD

十八、　其他