# マルコフ確率場を使ったタンパク質機能予測

秋葉寿光[1], 田口善弘[1,2]
[1]中央大学理工学研究科, [2]中央大学理工学研究所

近年マルコフ確率場の理論をタンパク質相互作用のデータに使うことで、相互作用のパターンから機能のアノテーションを行う研究が報告されている。今研究では、(1)パラメーターの算出方法を見直し、(2)直接的ではない第二近傍のタンパク質のペアを考慮したモデルを採用することで、機能予測の精度を上げることに成功した。モデルの有効性を確かめるために、交差検定を行った。タンパク質のデータとして出芽酵母を使い、機能を定義するために Gene Ontology(GO)を使った。タンパク質相互作用のデータは、Munich Information Center for Protein Sequences(MIPS)から入手したものを使用する。

# Markov random field models for protein function prediction

Hisamitsu Akiba[1], Y-h. Taguchi[1,2]
[1]Dept. phys., Chuo. Univ., akiba@phys.chuo-u.ac.jp
[2]Dept. Phys., Chuo. Univ., tag@granular.com

Predicting functions of unknown proteins is one of the most important problems in proteomics. It has been proposed to apply the theory of Markov random field (MRF) to infer proteins' functions using both protein-protein interaction data and the functional annotations of their interacted protein partners[1,2]. In this paper, we could improve accuracy of prediction by using several methods. First, we have adopted new parameter estimation method without using a logistic regression method. Second, we considered indirect relation protein-protein pairs. The efficiency of our prediction is measured by applying the leave-one-out cross validation procedure to a functional path matching scheme.

## 1. Introduction

In the present proteomics researches, inference of functions of unknown proteins is an important task. Gene expression profiles, the usage of phylogenetic profiles, and protein-protein interaction are typical approaches to this problem. As a computational approach, the application of Markov random field (MRF) to protein-protein interaction is proposed to predict protein functions and to annotate functions by considering interaction patterns. In this paper, we have successfully improved performance of inference by two procedures. First, estimation of parameters is reconsidered, i.e., we proposed to estimate them without using logistic regression used in the previous researches. Second, the second nearest neighbors of protein-protein interactions are considered to improve accuracy of function annotations. In order to see the efficiency of new models, we have applied cross validations to the results. MRF is used to protein-protein interactions in budding yeast and 82 selected Gene Ontology (GO) terms are predicted by model equations. Protein-protein interacion data is downloaded from Munich Information Center for Protein Sequences (MIPS)[3].

## 2. Materials and Methods

Protein-protein interaction data in budding yeast is downloaded from MIPS[3] and numeric values which are necessary to apply the following equations are estimated. In order that, the table of GO:ID for budding yeast is taken from Gene Ontology[4] data base and protein names are compared with the table. Hierarchical structures of Gene Ontology is considered by selecting 82 informative node (GO terms) including the list of all direct children (in more specific terms, one hierarchy level down ). Then fitness with model equations is estimated. 82 GO terms considered here are those include more than 30 proteins and any of those direct children do not include more than 30 proteins. This is the definition of informative node. Among the nearest neighbors of each protein, the number of proteins with informative nodes is computed. When $x$ is the binary vector of nodes in subnetwork to represent if each node is annotated by a GO term, probability is given by

$$P_r(x|\theta)=\frac{\exp(-U(x))}{Z(\theta)} \tag{1}$$

Here $Z(\theta)$ is defined as

$$Z(\theta)=\sum_x \exp(-U(x))$$

and $U(x)$ in eq. (1) is

$$U(x)=-\alpha N_1-\beta N_{10}-\gamma N_{11}-N_{00} \tag{2}$$

Here $\alpha,\beta,\gamma$ are parameters and $N_1$ is the number of proteins annotated with the considered GO term. $N_{10}$ is the number of pairs one of those is annotated by the GO term and another is not annotated. $N_{11}$ is the number of pairs both of those are annotated. $N_{00}$ is the number of pairs neither of those is annotated.

$$\log\frac{P_r(X_i=1|X_{[-i]},\theta)}{1-P_r(X_i=1|X_{[-i]},\theta)}=\alpha+(\beta-1)M_0^{(i)}+(\gamma-\beta)M_1^{(i)} \tag{3}$$

$P_r(X_i=1|X_{[-i]},\theta)$ in the left hand side of eq. (3) is the probability of annotation by the GO term for site *i* based upon the prior probability estimated by parameter $\theta$ and information excluding site *i*. Parameters are $\theta=(\alpha,\beta,\gamma)$. The left hand side of eq. (3) is the logarithm of the ration between the probabilities those protein *i* is annotated by the GO term or not. $M_0^{(i)}$ ( $M_1^{(i)}$ )is the number of proteins without (with) the GO term among the nearest neighboring proteins of protein *i*. Parameters $\alpha,\beta,\gamma$ are estimated by maximum likelihood method. Eq. (3) gives us the probability of GO annotation for the considered protein based upon prior probability of surrounded proteins.

### 2.1. New parameter estimation

Here we propose the method to estimate parameters without using maximum likelihood. First of all, eqs. (1) and (2) give us the possibilities when $N_{10,}N_{11,}N_{00,}N_1$ appear.

$$\frac{N_{10}}{N}=\frac{\exp(\alpha+\beta)}{Z}$$

$$\frac{N_{11}}{N}=\frac{\exp(2\alpha+\gamma)}{Z}$$

$$\frac{N_{00}}{N}=\frac{e}{Z}$$

$$\frac{N_1}{N'}=\frac{\exp(\alpha)}{Z}$$

Here $N$ is the total number of pairs of protein-protein interactions, $N'$ is the total number of proteins. By solving these four equations, we get $\alpha,\beta,\gamma$.

### 2.2. Indirect pair model

Until here, we consider sub network that consists of proteins which have relationships with the considered protein within protein-protein interaction data. Model equation is based upon this subnetwork. Here we extend the model so as to consider the effect of the second nearest neighboring proteins.

$$U(x)=-(\alpha N_1+\beta_1 N1_{11}+\gamma_1 N1_{10}+\delta_1 N1_{00}+\beta_2 N2_{11}+\gamma_2 N2_{10}+\delta_2 N2_{00}) \tag{4}$$

Here $NK_{ll'}$ is the total number of protein-protein interactions. $K=1$ means link to the nearest neighbors and $K=2$ means that to the second nearest neighbors. Substituting eq. (4) into eq. (1), we get the equation corresponding to eq. (3). Then using maximum likelihood, we estimate parameters and hence the probabilities of the considered protein is annotated by the specific GO term.

### 3.Results

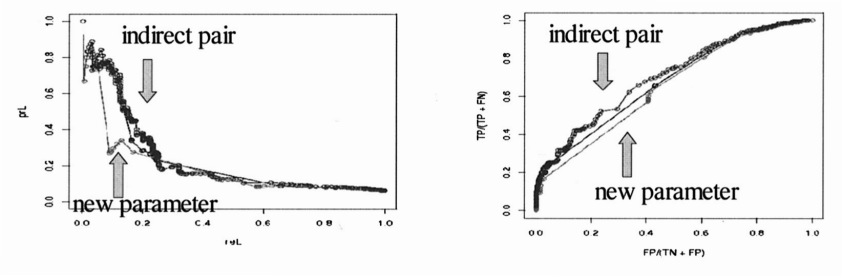
The performance measures are precision, recall, SN(sensitivity), FPR(false positive rate), and AUC score of ROC curve.

$$Precision=\frac{TP}{TP+FP} \qquad Recall=\frac{TP}{TP+FN} \qquad SN=\frac{TP}{TP+FN} \qquad FPR=\frac{FP}{TN+FP}$$

| ***GO:ID*** | ***logistic*** | ***new parameter*** | ***indirect pair*** |
|---|---|---|---|
| GO:0006412 | 0.67959551 | 0.654026 | 0.705483 |
| GO:0006974 | 0.7184796 | 0.7169281 | 0.773251 |
| GO:0007131 | 0.7209587 | 0.7051123 | 0.7976397 |

Table.1: AUC scores on GO:IDs. From the left column to the right one, conventional method (logistic), solving equations (new parameter), model considering the second neighbors (indirect pair).

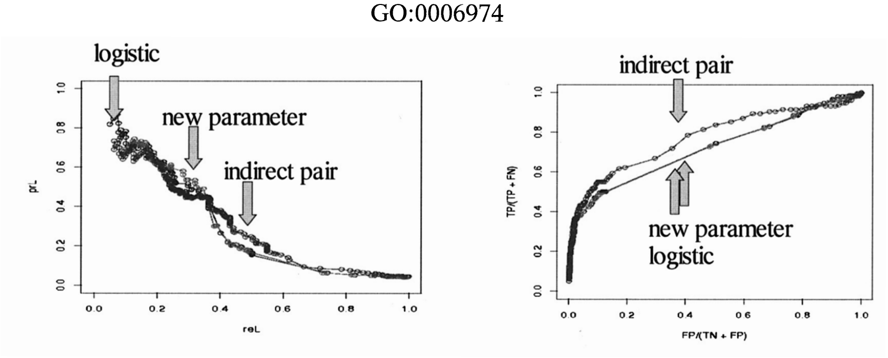
GO:0006412



(a)precision vs recall

(b)sensitivity vs false positive rate

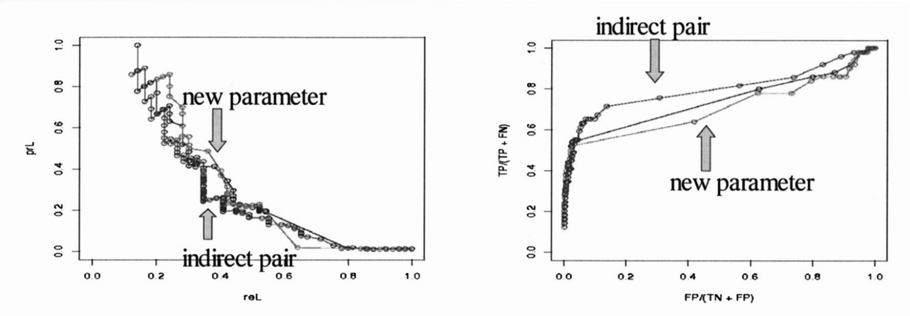
GO:0006974



(c)precision vs recall

(d)sensitivity vs false positive rate

GO:0007131



(e)precision vs recall (f)sensitivity vs false positive rate (ROC curve)

Figure 1:Performance for conventional method, solving equations, model considering indirect pair . Precision vs recall: (a),(c),(e), sensitivity vs false positive rate: (b),(d),(f).

### 4. Discussion

Reconsideration of parameter estimations and generating the new model considering the second nearest neighbor give us better performance of inference of functions annotated by Gene Ontology. Performance becomes better or worse dependent upon considered Gene Ontology. It is important to clarify the relationship between Gene Ontology and the model performance. Newly proposed method considering the second nearest neighbor does not take into account the difference if the nearest neighbor protein is annotated. The distinguish between the annotated and non-annotated nearest neighbor is important when we consider the second nearest neighbors.

### 5.Acknowledgement

This work has been partly supported by the Grant-in-Aid for Creative Scientific Research No.19500254 of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) from 2007 to 2008. We are grateful for their support.

### 6.Reference

[1]Deng, M., Zhang, K., Mehta, S., Chen, T. and Sun,F. (2002). Prediction of Protein Function Using Protein-protein Interaction Data. In Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB2002): 197-206.

[2]Deng M.,Tu Z.,Sun F.,Chen T(2003)Mapping Gene Ontology to proteins based on protein-protein interaction data. Bioinformatics. 2004 Apr 12;20(6):895-902. Epub 2004 Jan 29

[3]Munich Information Center for Protein Sequences(MIPS)
Mewes, H. W.MIPS - European node for protein sequence data.CODATA *Bulletin* 23: 62-63, 1991.

[4]The Gene Ontology(GO) Consortium(2000) Gene Ontology:tool for the unification of biology. Nat. Genet.,25,25-29.