行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※

※　　　基因調控網路之組合型重建 (3/3)　　　※

※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：■個別型計畫　　□整合型計畫

計畫編號：NSC　96－2221－E－009－042－

執行期間：96 年 8 月　1　日至　97 年 7 月 31 日

計畫主持人：胡毓志

共同主持人：

計畫參與人員：鄭家胤，王聖文

本成果報告包括以下應繳交之附件：

- □赴國外出差或研習心得報告一份
- □赴大陸地區出差或研習心得報告一份
- ■出席國際學術會議心得報告及發表之論文各一份
- □國際合作研究計畫國外研究報告書一份

執行單位：交通大學 資訊工程系

中　華　民　國　97 年　9　月　18　日

行政院國家科學委員會專題研究計畫成果報告

國科會專題研究計畫成果報告撰寫格式說明

Preparation of NSC Project Reports

計畫編號：NSC 96-2221-E-009-042

執行期限：96 年 8 月 1 日至 97 年 7 月 31 日

主持人： 胡毓志 交通大學 資訊工程系

計畫參與人員：鄭家胤、王聖文 交大 資工系

Abstract

了解基因表現的調控機制是分子生物學中一項重要挑戰，而基因調控網路的重建更是模擬此機制的基礎。在此報告裡，我們描述一種藉由整合基因表現與調控因子機率性的方法以重建基因網路。同時，為呈現其效能，我們將此方法測試於 27 個調控模組，為酵母菌中與細胞周期相關的 6 個轉錄因子及 15 個基因重建基因調控網路。

One of the major challenges in molecular biology is to understand the precise mechanism by which gene expression is regulated. Reconstruction of transcription networks is essential to modeling this mechanism. In this report, we describe a novel approach for building transcription networks from transcription modules by combining expression profile correlations with probabilistic element assessment. To demonstrate its performance, we systematically tested it on 27 transcription modules and reconstructed the transcription network for 6 transcription factors and 15 genes involved in the yeast cell cycle. The experimental results show that our combinatorial approach can better filter false positives to increase the selectivity in prediction of target genes. The regulatory control relationships described by the network reconstructed also mostly agree with those in earlier studies.

Introduction

Each cell is the product of specific gene expression programs specified by genomic sequences. These programs involve regulated transcription of thousands of genes (Lee, et al., 2002). The regulation of gene expression is very complex and often accomplished through the coordinated actions of multiple transcription factors (TFs) (Yuh et al., 1998; Halfon, et al., 2000; Fickett, et al., 2000). One way to understand the potential pathways that can be used by a cell to regulate global gene expression programs is to model the network of regulator-gene interactions.

As the advent of microarray technology, an enormous amount of gene-expression data from a variety of biological analyses has been generated (Spellman, et al., 1998; DeRisi, et al., 1997; Alon, et al., 1999). High-throughput and large-scale expression profiling is considered one of the most promising techniques for reconstructing genetic networks. The experimental results enable the global studies of gene regulation (van Berkum and Holstege, 2001). Inference of gene-expression regulatory mechanisms is rapidly becoming a major research topic in bioinformatics.

There has been much work on genetic regulatory networks, applying different network representations and inference strategies. For example, some is focused on conditional probability distribution in Bayesian networks, some derives Boolean functions for Boolean networks, and some is based on differential equations (D’haeseleer, et al., 2000; Akutsu, et al., 2000; Hartemink, et al., 2001; Chen, et al., 1999). More recent studies address the importance of the combinatorial nature of transcription. Using microarray data, they identify novel motif combinations and co-occurrence position preference (Pilpel, et al, 2001; Sudarsanam, et al, 2002). In addition, supervised learning is also adapted to reconstruct transcription networks (Soinov, et al., 2003).

Given the transcription factors and genes of interest, our goal here is to build a transcriptional regulatory network that can model the regulator-gene interactions. A transcription network can be decomposed into transcription modules (Wang, et al., 2002). Each module, related to specific cellular conditions or perturbations that control it, represents a functional unit consisting of a transcription factor, the target genes it regulates and the gene (or genes if the factor is a complex) that produces it. Based on a bottom-up strategy, we first identify transcription modules corresponding to particular gene expression profiles, and then we reconstruct a potential transcription network with the links among all the modules found.

The accuracy of a transcription network depends on correct transcription modules each of which defines the role of each gene in the module and its relation with the transcription factor. Building a correct module requires not only the identification of the conserved core of the DNA regulatory motif(s) in the upstream region recognized by a particular TF, but also the knowledge of the genes likely to be regulated by the TF and the gene(s) producing it. Most of the computational analyses of transcription factors and the corresponding genes have been concentrated on finding regulatory factor binding sites in the DNA sequences upstream of genes (Lawrence, et al, 1993; van Helden, et al., 1998; Hertz, et al., 1990). Despite many successful applications to predicting significant regulatory elements in groups of functionally related genes, the number of false positives of consensus pattern or matrix-based search in a large amount of sequence (e.g. genome size) is far from acceptable.

The development of large-scale expression monitoring and the availability of complete genome sequence allow the refinement of computational analysis. The combination of expression phenotype and sequence similarity has been suggested to increase the efficiency of *cis*-regulatory element prediction as well as to reduce the false positive rate of target gene search (Zhang, 1999; Wolfsberg, et al., 1999). However, few studies were systematically evaluated to determine if known elements were detected with a higher selectivity than in naïve searches.

In this report, we describe a novel approach for building transcription networks from transcription modules by combining expression profile correlations with probabilistic element assessment. We systematically evaluated the method across 121 transcript profile experiments with 27 different known factors. Furthermore, to demonstrate its effectiveness of reconstructing transcription networks, we applied our method to many cell cycle-related transcription factors and their target genes. We compared the predicted networks with those validated and published in literature.

Materials and Methods

Toward the Network of Transcription

By applying clustering techniques to the data from genome-wide expression monitoring studies, we can first obtain groups of genes according to the similarity of their expression levels. From each group we can then detect common *cis*-regulatory elements (Brazma, et al., 1998; Hu, et al., 2000; Fujibuchi, et al., 2001). Although similar gene expression behaviors can constrain the search space of interesting regulatory sequences, this type of approaches only takes into account the correlation between expression profile similarity and gene co-regulation. It neglects the potential interactions between the gene(s) that produces the transcription factor and those regulated. Unlike previous work, for a particular transcription factor, we integrate three different kinds of information to predict its regulatory sequences and build the transcription module. They include: (1) the transcription factor binding sites, (2) the expression profile similarity of potential target genes and (3) the correlation between the expression profile of the gene(s) that produces the factor and that of the target genes the factor regulates. By the synergy of various kinds of knowledge, we expect to better characterize the nature of transcriptional regulation mechanism, so as to improve the prediction of *cis*-regulatory sequences and ensure the quality of transcription modules.

We encode TF binding sites in the standard IUPAC/IUB code. For a particular transcription factor, we match its binding sites against the upstream region of entire genome. This provides the preliminary candidate *cis*-regulatory sequences. To filter spurious false positives from the pattern-based search, we develop a new metric that combines the probabilistic element assessment (PEA) with the p-value of F-test (PF) on regression. The PEA is a ranking of potential sites according to sequence similarity in the upstream regions of genes with similar expression profiles (Fujibuchi, et al., 2001). It evaluates the probability of element conservation in expression clusters based on the idea that a sequence pattern is a regulatory element (i.e. TF binding site) if observed more often than expected in a gene expression cluster. Assuming a binomial distribution, the PEA value is defined as follows:

$$P(k \geq x) = \sum_{i=x}^{N} \binom{N}{i} m^{i} (1-m)^{N-1}$$

where P is the probability of finding x or more genes whose upstream contains a specific regulatory element by chance, m is the expected probability of element occurrence and N is the total number of genes in the cluster. The value of m can be estimated from the fraction of genome that has the element. The lower the PEA is, the more significant the element is.

Besides applying PEA to measure the significance of a regulatory element with the correlation among the genes in an expression cluster co-regulated by a particular TF, we also try to model the association between the gene(s) that produces the TF and the TF's target genes. We perform a linear regression on their expression profiles, followed by an F-test, to evaluate the strength of the relationship. For a transcription factor, say F, produced by the genes, $x_1, x_2, \ldots, x_p$ (if F is a complex), we define the linear regression model as the following:

$$Y_i \approx \hat{Y}_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i}$$

where Y_i is the expression data of gene Y in the *ith* transcript profile experiment and $\hat{Y}_i$ is the corresponding estimated value of Y_i. We use $\hat{Y}_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i}$ as the estimating formula, where $x_{j,i}$ is the expression data of gene x_j in the *ith* transcript profile experiment, α is the constant, and $\beta_1 \ldots \beta_p$ are the regression coefficients. Our first step is to find the best fit of Y_i with $\hat{Y}_i$ by determining α and $\beta_1 \ldots \beta_p$, using the least square method, and then to verify whether all the genes, $x_1, x_2, \ldots, x_p$, taken together, significantly explain the observed Y_i. To test the significance, we form the following hypotheses:

H$_0$: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ null hypothesis: Y_i does not depend on the x_i's

H$_1$: at least one $\beta_i \neq 0$ alternative hypothesis: Y_i depends on at least one of the x_i's

We perform an F-test on the regression as a whole. If the null hypothesis is true, the ratio:

$$F = \frac{\dfrac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{p}}{\dfrac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-p-1}} \quad \text{where} \quad \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

has an F distribution with p numerator degrees of freedom and *n-p-1* denominator degrees of freedom. If the null hypothesis is false, then the F ratio tends to be larger than it is when the null hypothesis is true. We define $PF = prob(F_{p,n-p-1} > F)$, where $F_{p,n-p-1}$ is the F value corresponding to a significance level. Therefore, if PF (p-value of F-test) is significantly small, we reject H$_0$, and conclude that Y_i is associated with at least one of the x_i's. In our studies, a candidate gene is considered a target gene of a particular TF if: (1) the upstream region of the gene contains the *cis*-regulatory sequence of the TF, (2) the PEA value of the gene is significantly small and (3) the PF value (i.e. p-value of F-test) for regression analysis of the expression profiles is significantly small.

Take PDR3 for example. It recognizes the upstream binding site TCCGYGGA. To construct its transcription module, we first identify its potential target genes by searching the genome for genes whose upstream contains the binding site. For each potential target gene, we then compute the PEA value and conduct the regression analysis. Only those genes that have the specific binding site and significant PEA values as well as PF are selected to build the module. For instance, YBL005W, YDL011W and YDR406W all have the upstream binding site TCCGYGGA. Based on the 121 transcript profile experiments (DeRisi, et al.,1997; Eisen et al., 1998; Lashkari, et al., 1997; Chu, et al., 1998; Holstege, et al., 1998; Spellman, et al., 1998; Cho, et al., 1998; Jelinsky, et al., 1999), we further evaluate their PEA values and calculate the PF for the corresponding regression analyses. We use Figure 1 to illustrate the basic idea. The expression profile presents the mRNA level of each gene in the 121 transcript experiments. If the expression profiles of YBL005W, YDL011W and YDR406W are very similar, and can be well clustered together, we can obtain a significantly low PEA value for each of them. In addition, as YBL005W produces PDR3, we hypothesize that there exists a potential relationship between YBL005W and PDR3's target genes. This relationship is modeled by a linear regression. A significant p-value of F-test indicates the strong association. Combining all the analyses above, we may conclude that PDR3 probably regulates YBL005W, YDL011W and YDR406W. We can then build a potential transcription module for PDR3 as presented in Figure 2.

The same procedure can be applied to more transcription factors to build more transcription modules. The intra-module and inter-module interactions between the genes and the transcription factors form a network of transcriptional regulation as a result. A putative transcriptional regulatory network may look like that in Figure 3.

Data Preparation

Our current studies are focused on *Saccharomyces cerevisiae*. To keep the studies consistent with some earlier work (Fujibuchi, et al., 2001), we extracted 1000bp upstream of each of the 6194 yeast open reading frames, except seven sequences (YAL069W, YFL067W, YFL068W, YJR162C, YKL225W, YMR326C and YNR077C). They have a shorter upstream region owing to their occurrence close to a chromosome end.

We used the same expression experiment data as Fujibuchi's, since all the data are available in public databases and had been reported in literature (Fujibuchi, et al., 2001). Some of the data represented the time courses in time series-based experiments, and some were obtained from multiple (or single) experiments under various control conditions. The expression datasets are summarized in Table 1. We described each yeast gene as a vector of 121 experimental data elements. Each element was stored as the ratio of expression levels in two states, experimental and reference. The only exception is the data from Cho et al. (Cho et al., 1998). Since only a single value was available in their original work, we used that value directly as one element. The data elements were further processed in the same way as in the work of Spellman et al. (Spellman et al., 1998) to normalize the sum of all values within a specific experiment to zero. We applied a Pearson correlation coefficient-based hierarchical clustering algorithm (Eisen, et al., 1998) to the normalized expression data and derived a clustering result similar to that of Fujibuchi's (Fujibuchi, et al., 2001). The clustering result and the normalized expression data were the basis of computing PEA values and performing regression analyses.

Results

Identification of Transcription Modules

As the accuracy of a transcription network depends on the correctness of its transcription modules, before verifying our method can reconstruct meaningful transcription networks, we demonstrate its performance of building significant transcription modules.

A transcription module is composed of a particular transcription factor and its target genes. Although several techniques have been proposed to locate the target genes of a transcription factor, none of them was systematically evaluated to determine if known regulatory elements were better identified than by naïve searches (Quandt, et al., 1995; Lavorgna, et al., 1999; Zhang, 1999; Wolfsberg, et al., 1999). , To justify the feasibility of our new approach, we compared our method with PROSPECT (Fujibuchi, et al., 2001), which had been systematically evaluated. Like PROSPECT, we selected SCPD as the dataset for evaluation. We merged the information of recognition sites from SCPD (Zhu and Zhang, 1999) and TRANSFAC (Wingender, et al., 1996). After removing the sites without sensible consensus and disregarding the regulators for which the gene(s) producing them is unknown, we selected 27 regulatory elements in our studies. We also used the same metric as in PROSPECT, the selectivity ratio, for comparison. For a particular method, the selectivity is the fraction of correctly predicted elements out of all elements predicted, and the selectivity ratio is defined as the ratio of its selectivity to that of a naïve pattern match method. In addition to selectivity, sensitivity is another important evaluation criterion. Despite the lack of full knowledge of true target genes, we simply define sensitivity as the fraction of correctly predicted elements out of all elements annotated in SCPD, assuming SCPD is complete.

One important hypothesis behind our method is that there may exist potential relationships between the gene(s) producing a transcription factor and those regulated by this particular factor. Our experiments justified the hypothesis by showing that the target genes could be predicted by exploring the relationships through regression analysis. A lower PF value of the regression analysis between a transcription factor and genes suggests a stronger relationship between the transcription factor and these genes. To examine the effect of PF values on the selectivity ratio, we varied the PF threshold when testing the 27 elements. The results showed that for 16 elements, the selectivity ratio increased up to two or higher along the decrease of PF threshold. These 16 elements are ACE2, BAS1, BAS2, GCN4, GCR1, HAP1, HSTF, LEU3, MIG1, PDR3, PUT3, RAP1, REB1, SBF, SFF and STE12. The selectivity ratio fluctuated between one and two for nine elements, ABF1, ADR1, GAL4, MATα1, MATα2, MCM1, MBF, PHO4 and repressor of CAR1. Only for SWI5 and TBP did the decrease of PF threshold make the selectivity ratio worse instead. We show some of the results in Figure 4. Our empirical studies showed that when PF=0.03 and PEA=0.4, regression analysis and PROSPECT have the best overall performance respectively over the 27 elements. We compared their results with those of naïve pattern searches. The results are summarized in Table 2. For selectivity ratio, regression analysis outperformed PROSPECT in 16 elements but lost in 11 elements; as for sensitivity, PROSPECT did better in 14 elements and tied in four elements. The experimental results suggest that there indeed exists the association between some transcription factors and the target genes, and this relationship can be characterized by regression analysis.

Since neither of regression analysis and PROSPECT outperformed the other in the prediction of all the 27 elements, we combined both approaches to benefit from the synergy. We consider a gene as a target gene of a particular transcription factor only if its PEA value and PF value are both smaller than some significance thresholds. We use both PEA and PF if available as the criteria to filter out false positives. To verify the performance of the combinatorial approach, we compared the performance of applying PEA or PF alone with that of using both. We carried out a systematic evaluation, using various values of PEA and PF (between 0.4 and 10^{-6}), over the 27 regulatory elements. We found that the combinatorial approach obtained higher selectivity ratio than either one alone in the prediction of each element. It proved the synergy of PEA and PF. By applying PEA and PF together, our method identifies the target genes of a particular transcription factor to build the transcription module.

Reconstruction of Transcription Networks

Given the transcription factors and the genes of interest, our goal is to reconstruct a transcription network that can model the interactions among them. In this specific case, we applied our method to the transcription factors and genes that are involved in the yeast cell cycle. We chose six regulators (MCM1, ACE2, SWI5, SBF, MBF and SFF) and fifteen genes (CLB1, CLB2, SWI5, ACE2, CDC5, CLN3, SWI4, FAR1, RME1, SIC1, CDC6, CLN1, CLN2, CLB5 and CLB6) in our studies. They play an important role in the yeast cell cycle (Mendenhall et al., 1998). Their functions are described in Table 3. We set the PF threshold at 0.03 and the PEA threshold at 0.65. We reconstructed the network of transcriptional regulation as shown in Figure 5. Comparing Figure 5 to Table 3, we found our method correctly identified that transcription factor SFF regulates ACE2, CLB1, CLB2 and SWI5, ACE2 regulates RME1, MCM1 regulates SWI4, and SBF regulates CLN1 as well as CLN2. In addition, the network shows that SFF indirectly regulates CDC5 via SWI5.

Discussion

The reconstruction of transcriptional regulatory networks is essential to understanding how regulators and genes interact. Based on the hypothesis that co-regulated genes have similar expression profiles and genes producing transcription factors have strong correlation with regulated genes, we combine probabilistic element assessment, regression analysis and binding site information to build transcription modules in a network.

Our combinatorial approach has several advantages. First, each metric covers different kinds of background knowledge. Because we exploit more information to identify transcription modules, we can better filter false positives. Second, these metrics complement each other by characterizing different biological activities, e.g. similar expression profiles among co-regulated genes and the associations between regulators and target genes. Third, our combinatorial approach is more robust. If one metric is not applicable, our system is still functional with the other metrics available. For example, in case the binding site sequences are unknown, we can still identify reasonable transcription modules (with more false positives though), applying only regression analysis. The intra-module and inter-module links connect all the components (genes and regulators) to form a transcription network. The links indicate either direct or indirect regulatory control. Our experimental results show the relationships that agree with those in previous studies (Mendenhall et al., 1998).

The current method can be further improved in several directions. Although the incorporation of PF values can generally increase selectivity ratio, our experiments showed some cases in which the use of PF thresholds could be harmful. Some true positives were mistakenly filtered out. Two possible causes of the mistake are: (1) the correlation is only implied in part of the expression profiles, so the irrelevant expression data may mislead the regression analysis; (2) the correlation cannot be accurately characterized by linear regression. One feasible solution to the first problem is to partition the whole expression profile into segments based on types of transcript experiments or on time intervals. Regression analysis is done on segments separately to reduce the noise caused by irrelevant expression data. As for the second problem, we require more domain knowledge to revise our hypothesis about the association between genes producing transcription factors and those regulated. Besides, high-order regression analyses may be desirable.

References

Akutsu, T., Miyano, S. and Kuhara, S. "Algorithms for inferring qualitative models of biological networks", in Proceedings of Pacific Symposium of Biocomputing, p293-304, 2000.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proceedings of National Academic Science, USA*, 96, p6745-6750, 1999.

Brazma, A., Jonassen, I., Vilo, J. and Ukkonen, E. "Predicting gene regulatory elements in silico on a genomic scale", *Genome Research*, 8, p1202-1215, 1998.

Chen, T., He, H.L. and Church, G.M. "Modeling gene expression with differential equations", in Proceedings of Pacific Symposium of Biocomputing, p29-40, 1999.

Cho, R.J., Campbell, H.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Lansman, D., Lockhart, D.J. and Davis, R.W. "A genome-wide transcriptional analysis of the mitotic cell cycle", *Molecular Cell*, 2, p65-73, 1998.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. "The transcriptional program of sporulation in budding yeast", *Science*, 282, p699-705, 1998.

DeRisi, J.L., Iyer, V.R. and Brown, P.O. "Exploring the Metabolic and Genetic Control of Gene Expression on A Genomic Scale", *Science*, 278, p680-686, 1997.

D'haeseleer, P., Liang, S. and Somogyi, R. "Genetic network inference: from co-expression clustering to reverse engineering", *Bioinformatics*, 16, p707-726, 2000.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. "Cluster analysis and display of genome- wide expression patterns", *Proceedings of National Academic Science, USA*, 95, p14863-14868, 1998.

Fickett, J.W. and Wasserman, W.W. "Discovery and modeling of transcriptional regulatory regions”, *Current Opinions in Biotechnology*, 11, p19-24, 2000.

Fujibuchi, W., Anderson, J. and Landsman, D. “PROSPECT improves *cis*-regulatory element prediction by integrating expression profile data with consensus pattern searches”, *Nucleic Acids Research*, 29, p3988-3996, 2001.

Halfon, M.S. et al. “Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors”, *Cell*, 103, p63-74, 2000.

Hartemink, A.J., Gifford, D. K., Jaakkola, T.S. and Young, R.A. “Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks”, in Proceedings of Pacific Symposium of Biocomputing, p422-433, 2001.

Hertz, G., Hartzell III, G. and Stormo, G. (1990) Identification of Consensus Patterns in Unaligned DNA Sequences Known to be Functionally Related. *Comput. Appl. Biosci.*, 6, p81-92.

Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S. and Young, R.A. “Dissecting the regulatory circuitry of a eukaryotic genome”, *Cell*, 95, p717-728, 1998.

Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. “Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*”, *Journal of molecular Biology*, 296, p1205-1214, 2000.

Hu, Y., Sandmeyer, S., McLaughlin, C. and Kibler, D. “Combinatorial Motif Analysis and Hypothesis Generation on A Genomic Scale”, *Bioinformatics*, 16, p222-232, 2000.

Jelinsky, S.A. and Samson, L.D. “Global response of *Saccharomyces cerevisiae* to an alkylating agent”, *Proceedings of National Academic Science, USA*, 96, p1486-1491, 1999.

Lavorgna, G., Guffanti, A., Borsani, G., Bllabio, A. and Boncinelli, E. “TargetFinder: searching annotated sequence databases for target genes of transcription factors”, *Bioinformatics*, 15, p172-173, 1999.

Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O. and Davis, R.W. “Yeast microarrays for genome wide parallel genetic and gene expression analysis”, *Proceedings of National Academic Science, USA*, 94, p13057-13062, 1997.

Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A. and Wootton, J. (1993) Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignments. *Science*, 262, 208-214.

Lee, T. et al., “Transcriptional regulatory networks in *Saccharomyces cerevisiae*”, *Science,* 298, p799-804, 2002.

Mendenhall, M.D. and Hodge, A.E. “Regulation of cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast *Saccharomyces cerevisiae*”, *Microbiology Molecular Biology Review*, 62, p1191-1243, 1998.

Pilpel, Y., Sudarsanam, P. and Church, G.M. “Identifying regulatory networks by combinatorial analysis of promoter elements”, *Nature Genetics*, 29, p153-159, 2001.

Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. “MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data”, *Nucleic Acids Research*, 23, p4878-4884, 1995.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Bostein, D., and Futcher, B. “Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization”, *Molecular Biology of the Cell*, 9, p3273-3297, 1998.

Soinov, L.A., Krestyaninova, M.A. and Brazma, A. “Toward reconstruction of gene networks from expression data by supervised learning”, Genome Biology, 4:R6, 2003. (electronic version can be found at http://genomebiology.com/2003/4/l/R6)

Sudarsanam, P., Pilpel, Y. and Church, G.M. “Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*”, *Genome Research*, p1723-1730, 2002.

van Berkum N.L. and Holstege, F. C. “DNA microarrays: raising the profile”, *Current Opinions in Biotechnology*, 12, p48-52, 2001.

van Helden, J., Andre, B, and Collado-Vides, J. (1998) Extracting Regulatory Sites from the

Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies. *Journal of Molecular Biology*, 281, p827-842.

Wang, W., Cherry, J. M., Botstein, D. and Li, Hao "A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*", *Proceedings of National Academic Science, USA*, 24, p16893-16898, 2002.

Wingender, E., Dietze, P., Karas, H. and Knuppel, R. "TRANSFACT: a database on transcription factors and their DNA binding sites", *Nucleic Acids Research*, 24, p238-241, 1996.

Wolfsberg, T.G., Gabrielian, A.E., Campbell, M.J., Cho, R.J., Spouge, J.L. and Landsman, D. "Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*", *Genome Research*, 9, p775-792, 1999.

Yuh, C.H., Bolouri, H. and Davidson, E.H. "Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene", *Science*, 279, p1896-1902, 1998.

Zhang, M.Q. "Promoter analysis of co-regulated genes in the yeast genome", *Computational Chemistry*, 23, p233-250, 1999.

Zhu, J. and Zhang, M.Q. "SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*", *Bioinformatics*, 15, p607-611, 1999.

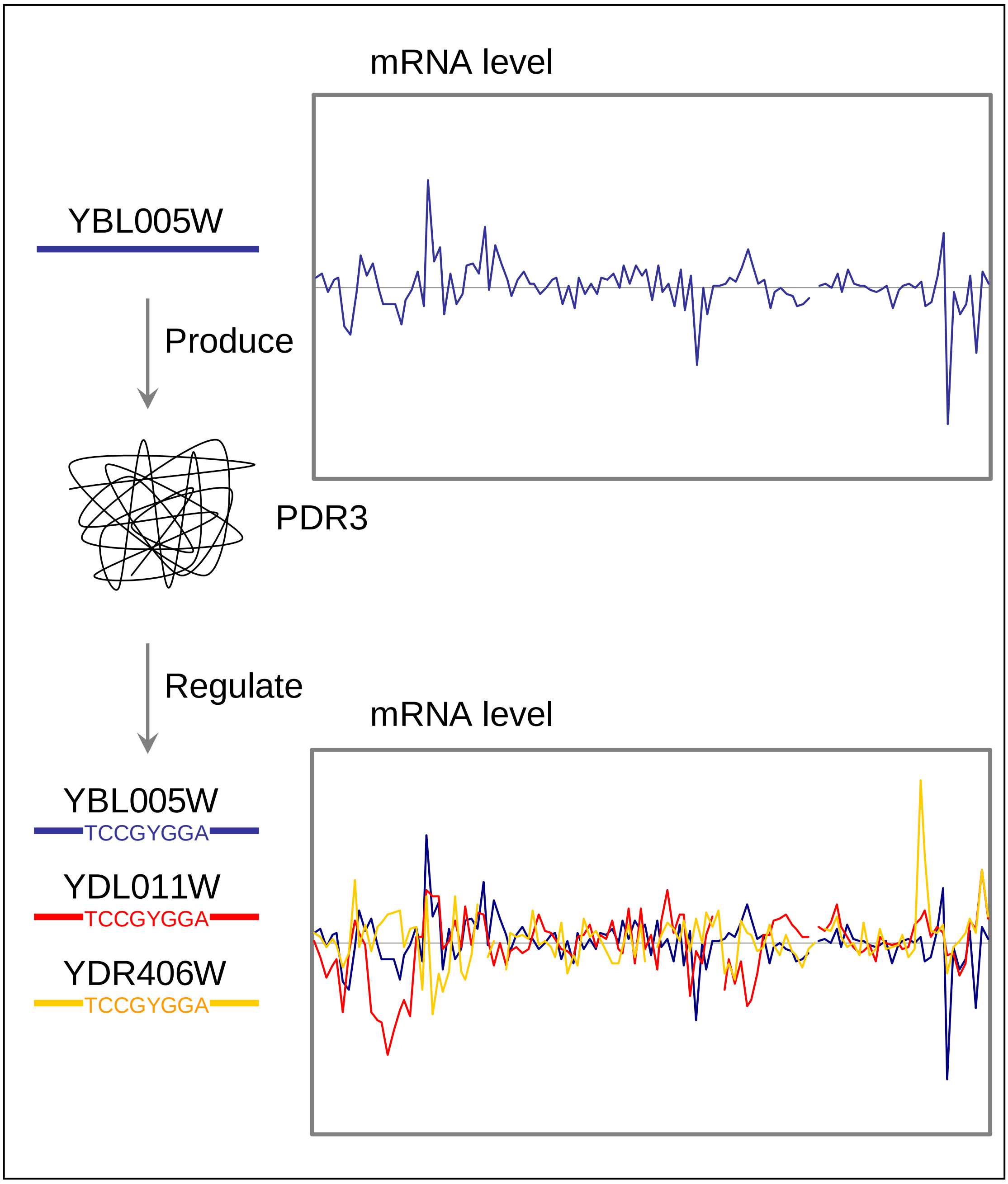


Figure 1. Synergy of binding sites, co-expression of target genes via a TF and correlation between a TF and its target genes. PDR3 regulates YBL005W, YDL011W and YDR406W, each of which contains the binding site TCCGYGGA upstream of the gene. The similarity of their mRNA levels leads to a low PEA value and suggests that the genes are co-regulated via PDR3. The correlation between PDR3 and its target genes is reflected by the relationship between YBL005W's mRNA level (which produces PDR3) and those of YDL011W and YDR406W. The relationship can be modeled by regression analysis and its significance is measured by F-test.

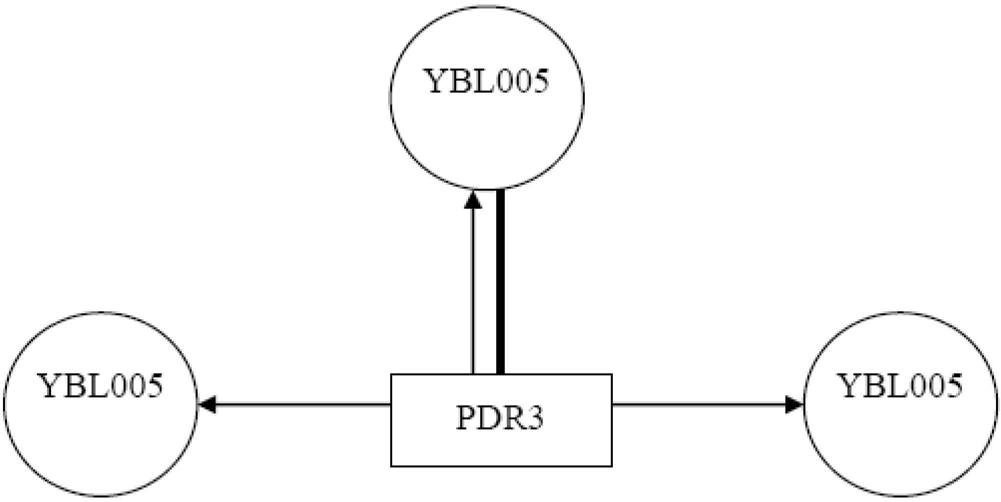


Figure 2. A transcription module of PDR3. The circles pointed by an arrow are the target genes of PDR3. The undirected edge between YBL005W and PDR3 indicates that YBL005W produces PDR3.

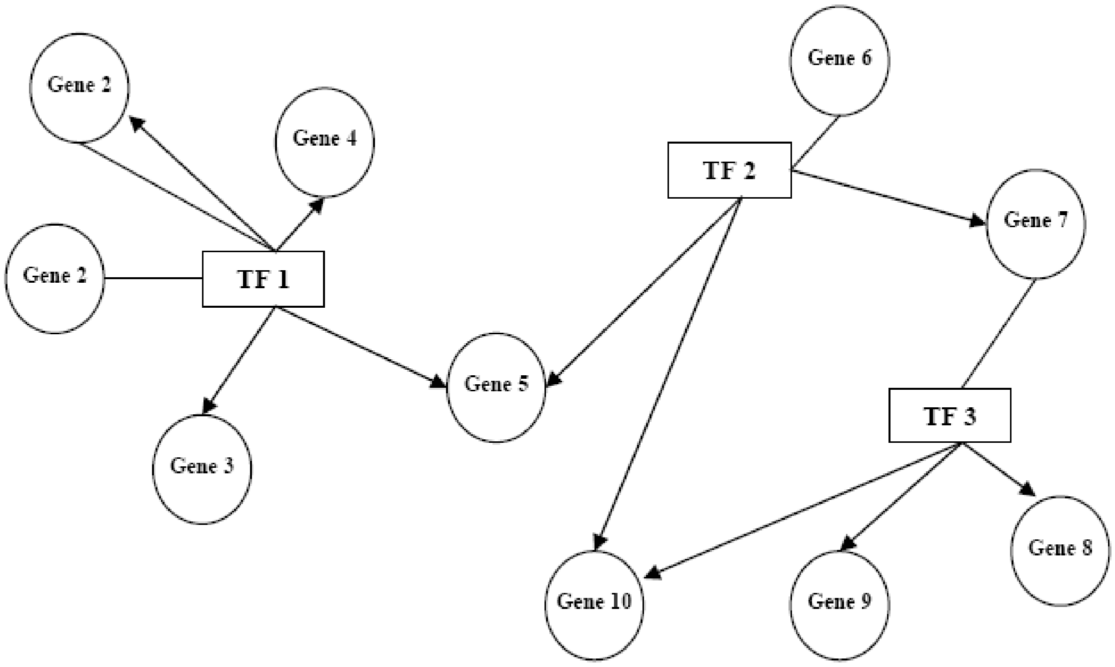


Figure 3. An illustration of a transcription network. Circles and squares stand for genes and transcription factors respectively. Undirected edges represent the production relationships between genes and transcription factors. For example, Gene 1 and 2 produce TF 1. Arrows, on the other hand, indicate the regulation relationships between transcription factors and genes. For example, TF 2 regulates gene 5, 7 and 10.

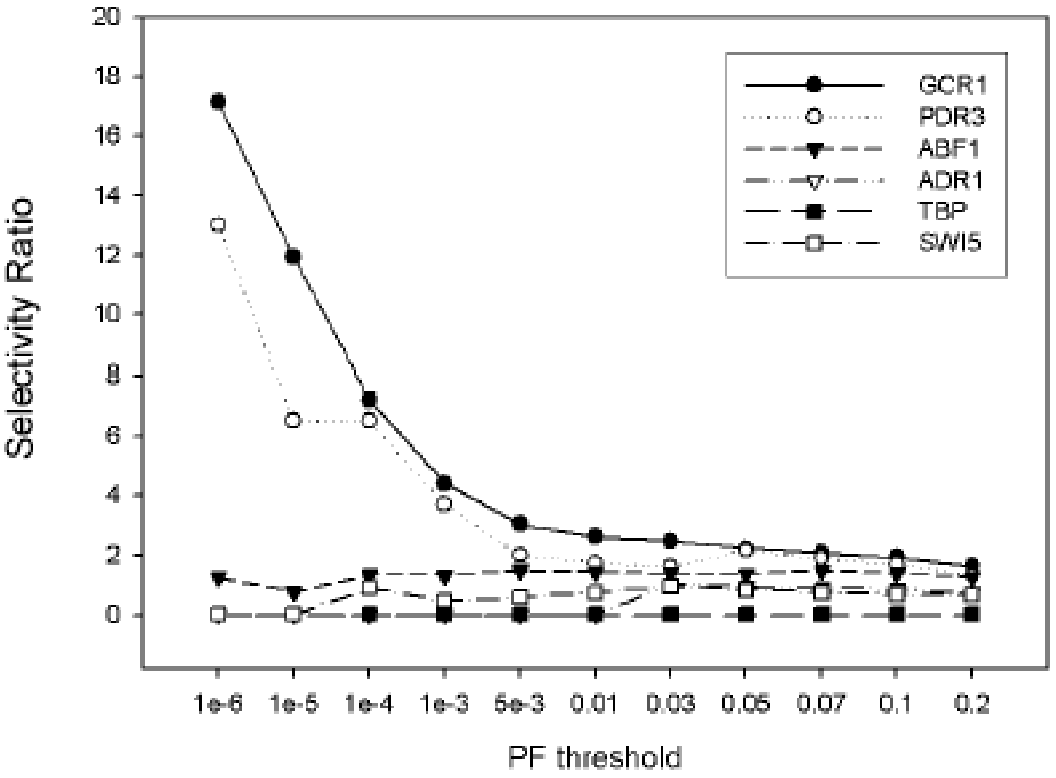


Figure 4. The effect of PF threshold on selectivity ratio. The decrease of PF threshold caused the selectivity ratio of GCR1 and PDR3 to increase dramatically. The selectivity ratio of ABF1 and ADR1 varied between one and two. On the other hand, as PF threshold decreased, the selectivity ratio of TBP and SWI5 got lower.

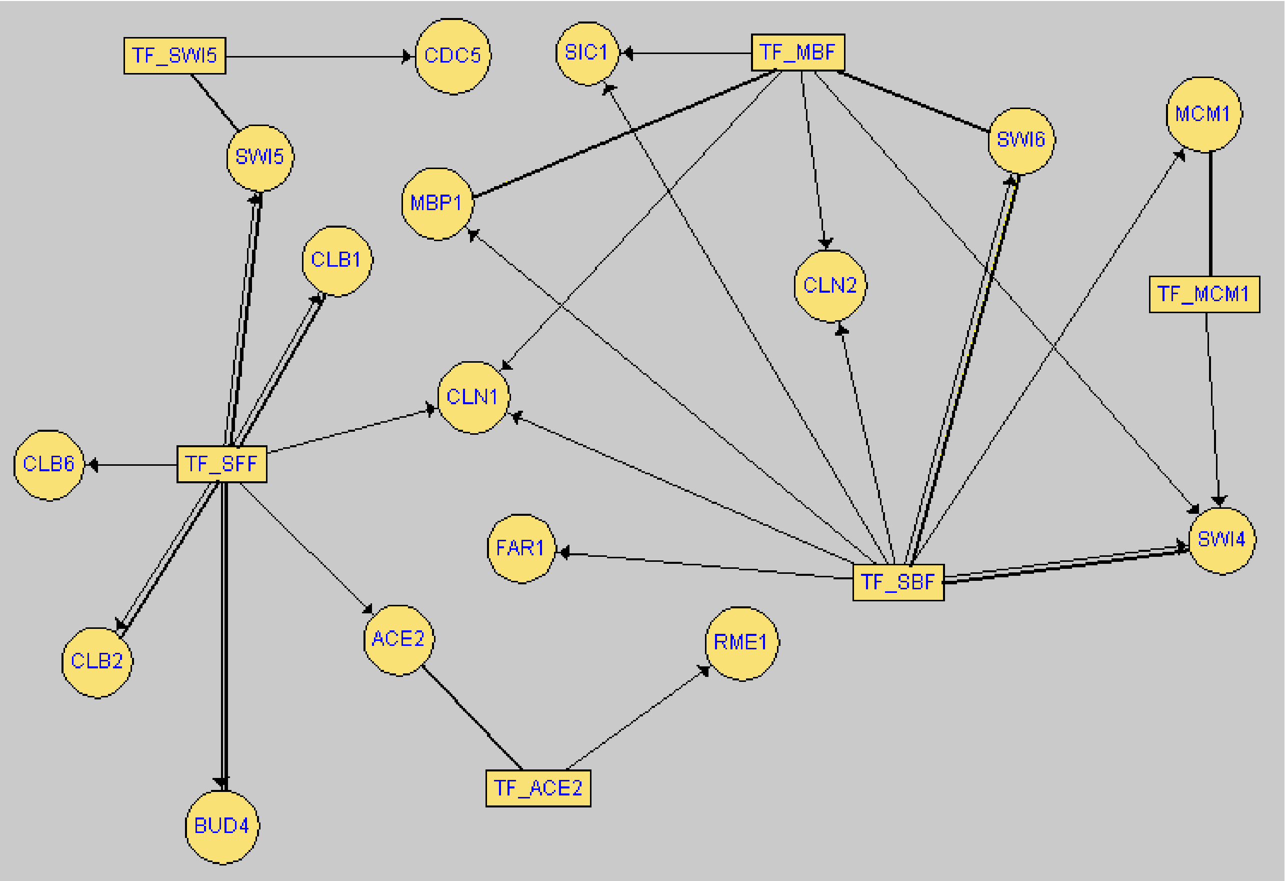


Figure 5. The transcription network of several transcription factors and genes involved in the yeast cell cycle. This network correctly presents several regulatory relations. For example, transcription factor SFF directly regulates ACE2, CLB1, CLB2 and SWI5, ACE2 regulates RME1, MCM1 regulates SWI4, and SBF regulates CLN1 as well as CLN2. There are also some other interactions that need to be further studied, e.g., the network shows that SFF indirectly regulates CDC5 via SWI5.

Reference	Dataset Description	Type	Experiments
DeRisi et al. 1997	Diauxic shift, repressor TUP deletion, activator YAP1 overexpression	Time series, cDNA microarray	9
Eisen et al. 1998; Lashkari et al. 1997	Heat shock, DTT shock, cold shock	Time series, cDNA microarray	14
Chu et al. 1998	Sporulation, sporulation ndt80 knockout	Time series, cDNA microarray	9
Holstege et al. 1998	Transcription factor mutant, SAGA chromatin modification complex mutant	Multiple experiments, oligonucleotide chip	11
Spellman et al. 1998	Cell cycle -facotr arrest, cell cycle elutriation, cdc15 arrest	Time series, cDNA microarray	60
Cho et al. 1998	cdc28 arrest	Time series, oligonucleotide chip	17
Jelinsky et al. 1999	Alkylating agents	Single experiment, oigonucleotide chip	1
Total			121

Table 1. Summary of gene expression profiles datasets. The third column indicates the type of the biochip used and whether the data are time courses collected at various time points or expression levels obtained from multiple/single transcript experiments.

Regulatory element	Naïve Pattern Search			Regression Analysis (PF=0.03)				PROSPECT (PEA=0.4)			
	TP	P	selectivity	TP	P	selectivity ratio	sensitivity	TP	P	selectivity ratio	sensitivity
ABF1	19	2974	0.006389	11	1237	1.391907	0.578947	13	881	2.309696	0.684211
ACE2	1	1861	0.000537	1	312	5.964744	1.000000	1	1410	1.319858	1.000000
ADR1	2	5018	0.000399	1	2413	1.039785	0.500000	0	784	0.000000	0.000000
BAS1	4	1602	0.002497	3	760	1.580921	0.750000	3	1143	1.051181	0.750000
BAS2	2	5861	0.000341	1	2779	1.054516	0.500000	0	1276	0.000000	0.000000
GAL4	6	349	0.017192	2	102	1.140523	0.333333	6	312	1.118590	1.000000
GCN4	9	6193	0.001453	5	2477	1.389001	0.555556	2	1158	1.188448	0.222222
GCR1	6	6016	0.000997	6	2425	2.480825	1.000000	6	1122	5.361854	1.000000
HAP1	4	61	0.065574	2	19	1.605263	0.500000	4	58	1.051724	1.000000
HSTF	6	5225	0.001148	5	1758	2.476773	0.833333	2	1036	1.681145	0.333333
LEU3	2	37	0.054054	1	14	1.321429	0.500000	1	34	0.544118	0.500000
MATα1	3	2035	0.001474	1	669	1.013951	0.333333	3	1624	1.253079	1.000000
MATα2	7	2178	0.003214	3	681	1.370673	0.428571	7	1794	1.214047	1.000000
MBF	6	1677	0.003578	4	801	1.395755	0.666667	6	1188	1.411616	1.000000
MCM1	25	2498	0.010008	3	441	0.679728	0.120000	22	1748	1.257574	0.880000
MIG1	7	652	0.010736	2	281	0.662938	0.285714	6	539	1.036841	0.857143
PDR3	7	182	0.038462	3	49	1.591837	0.428571	7	167	1.089820	1.000000
PHO4	4	2209	0.001811	1	711	0.776723	0.250000	4	1806	1.223145	1.000000
PUT3	1	282	0.003546	0	146	0.000000	0.000000	1	230	1.226087	1.000000
RAP1	15	2035	0.007371	12	964	1.688797	0.800000	11	1585	0.941535	0.733333
REB1	12	1440	0.008333	5	635	0.944882	0.416667	11	1261	1.046788	0.916667
Repressor of CAR1	13	471	0.027601	8	251	1.154766	0.615385	11	391	1.019280	0.846154
SBF	3	3692	0.000813	3	1245	2.965462	1.000000	2	1011	2.434553	0.666667
SFF	3	761	0.003942	3	469	1.622601	1.000000	2	613	0.827624	0.666667
STE12	4	4790	0.000835	1	1734	0.690600	0.250000	0	959	0.000000	0.000000
SWI5	1	4676	0.000214	0	534	0.000000	0.000000	1	1257	3.719968	1.000000
TBP	16	4944	0.003236	5	1629	0.948435	0.312500	3	692	1.339595	0.187500

Table 2. Results of naïve pattern search, regression analysis and PROSPECT. TP is the number of true positive regulatory elements in yeast genome. P is the number of elements predicted.

Cell Cycle	TF	Target genes	Functions
M/G1, Early G1	MCM1	CLN3	Cyclin activator of CDC28 in G1.
		SWI4	DNA binding component of SBF transcription factor. Important for Start-specific expression of CLN1 and CLN2.
		FAR1	CKI specific for CDC28-CLN complexes.
	ACE2	RME1	Positive factor in CLN2 expression. Negatively regulates early sporulation-specific genes.
		SIC1	CKI specific for CDC28-CLB complexes.
	SWI5	CDC6	Required for DNA replication. Inhibitor of CLB-CDC28 complexes.
Start (late G1)	SBF	CLN1	Cyclin activator of CDC28 at Start .
		CLN2	Cyclin activator of CDC28 at Start .
	MBF	CLB5	Cyclin activator of CDC28 at Start .
		CLB6	Cyclin activator of CDC28 at Start .
G2	SFF/MCM1	CLB1	Cyclin activator of CDC28 at G2/M .
		CLB2	Cyclin activator of CDC28 at G2/M .
		SWI5	Transcription factor important for expression of SIC1, CDC6, and RME1.
		ACE2	Transcriptional activator of SIC1 and RME1.
		CDC5	Protein kinase of the "plol" family. Activator of the APC.

Table 3. Summary of yeast cell cycle-related transcription factors and genes.

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

附件一

97 年 9 月 18 日

報告人姓名	胡 毓 志	服務機構及職稱	交通大學資訊工程系 副教授
時間 會議 地點	07/14/2008-07/17/2008 Las Vegas, U.S.A.	本會核定補助文號	NSC 96-2221-E-009-042-
會議名稱	（中文）Biocomp 生物資訊暨計算生物學國際研討會 （英文）Biocomp 2008		
發表論文題目	1.（中文）利用蛋白質結構字元表述蛋白質區間特性 （英文）Using Protein Structural Alphabet to Characterize Local Structure Features		

一、參加會議經過

於 07/13 辦理註冊報到，隔日隨即參加開幕演說，於 07/14-07/17 期間，參加與會學者之論文發表，並與多位國外學者討論相關研究議題。會議中不乏中國大陸籍學者之論文，對於我國內生物資訊的發展，應可產生良性刺激，提供非常多的助益與新的發展方向。

二、與會心得

根據議程中部分美國研究學者所述，由於經濟壓力上升，美國 NIH 已將研究主軸放在 translational research，希望藉由在實驗室的研究成果實際應用於人類醫學。本次參加人數及國家眾多，其研究領域更包括計算機科學、醫學、生物學等之應用，藉由討論及論文發表，獲得寶貴經驗，對於未來研究提供了新的方向。其中更結識他國友人，經由研討，可明白其他國家的發展經驗。從這次與會學習的經驗，我們可以得知國外研究之重點，作為我國在生物科技的發展依據。

三、考察參觀活動(無是項活動者省略)

無

四、建議

生物科技是目前國內新興研究發展之重要產業，懇請國科會及相關單位，能多支持與獎勵國內學者多參與此類國際研討會，除了增加我國在國際相關領域的能見度，同時，提供相互學習之機會。此外，建議由國科會主導，召集國內各大學與民間企業支援，以召開國際性生物資訊與相關科技研討會，邀請國內外學者共同參與，這是直接提昇我國在生技發展地位的最有效做法。

五、攜回資料名稱及內容

The Proceedings of Biocomp2008

Using Protein Structural Alphabet to Characterize Local Structure Features

Shih-Yen Ku[12] **and Yuh-Jyh Hu**[13]
[1]Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
[2]MIB Program at Institute of Statistical Science, Academia Sinica, Taipei, Taipei
[3]Institute of Biomedical Engineering, National Chiao Tung University, Hsinchu, Taiwan

Abstract - **As the number of available 3D protein structures increases rapidly, a wider variety of studies can be conducted more efficiently, among which is the design of protein structural alphabet. With the structural alphabet, not only can we describe the global folding structure of a protein as a 1D sequence, but we can also characterize local structures in proteins. Previously, we applied a combinatorial approach to protein structural alphabet design. In our previous work, we verified the usefulness of our structural alphabet by demonstrating the competitive accuracy in protein alignment, compared with alphabets. Here we took a further step by applying motif finding tools to our alphabet with the aim to characterize protein structure local features. Two structure domains, TIM and EGF, were used to evaluate the performance of our structural alphabet. Our method successfully recovered their sub-domains as common motifs in our structural alphabet.**

Keywords: protein structure, structural alphabet, motifs

Introduction

As all proteins have a certain degree of structural similarities to other proteins, and they probably share a common ancestor in evolution. Based on evolutionary relationships and the principles governing the 3D structures, a protein structure hierarchy, SCOP, was constructed mainly by visual inspection with the assistance of various automatic tools to compare protein structures. The original aim of SCOP was to serve as a tool for understanding protein evolution through the relationships between sequences and structures [1].

The conservation in local active sites may reflect biological meanings, and their structural patterns can be used to predict protein functions [2], e.g., the binding sites for metal-binding proteins [3]. The conserved local structural features can be identified in various ways and described in different representations. For example, some have attempted to investigate the relationships between local sequences and structures by identifying common structural motifs first, then characterizing amino acid preferences [4-6]. Others instead have adopted the inverse approach by examining structural correlates from recurring sequence patterns found to obtain sequence-structure motifs [7,8].

Unlike those works above on correlations between protein local structures and sequence patterns, we first convert protein 3D structures into 1D structural alphabet letters, and then identify and represent conserved local features as 1D structural alphabet sequence motifs. Besides, our goal is to mine the protein families for conserved local characteristics rather than to predict 3D structures of novel proteins as those studies mentioned above. There are several advantages of 1D structural alphabet over 3D co-ordinates representations. First, 1D representation of protein structures is more efficient in comparison and more economical in storage. Second, many previously designed and widely used 1D sequence alignment tools can be directly applied to protein structures as well as sequences. Third, conserved protein local structural features can be described as 1D sequence motifs and be identified by various well-developed sequence motif-finding tools. Four, this type of 1D-based approaches can serve as a pre-processor to filter out remotely related or irrelevant proteins before we apply other more accurate but more computationally intensive structure analysis tool.

Previous analysis of protein structures has shown the importance of repetitive secondary structures, in particular, α-helix and β-sheet. Together with variable coils, they constituted a basic standard 3-letter structural alphabet. In spite of the increase in predictive accuracy, the approximation of 3D structures with only a 3-letter alphabet is apparently too crude for the more refined 3D reconstruction [9-13]. Various more complex structural alphabets have been developed by taking into account the heterogeneity of backbone protein structures through sets of small protein fragments frequently observed in different protein structure databases [14-21]. Unlike most other works, we developed a multi-strategy method for structural alphabet design, which combined self-organizing maps, minimum spanning tree algorithm and k-means algorithm [22]. The performance of our alphabet was demonstrated by the competitive accuracy in all-alpha protein search within SCOP using the standard 1D sequence alignment tool, FASTA [23].

In this paper, we introduced an improved version of our alphabet design pipeline, to which we added a substitution matrix self-trainer. The substitution matrix used in aligning proteins represented by structural alphabets affects the accuracy of alignment. In our earlier work, we applied the identity matrix in the alignment [22]. Though the preliminary results successfully demonstrated the feasibility of our alphabet, yet a more appropriate matrix will further improve its applicability. The substitution matrix is a crucial factor in the successful application of 1D sequence alignment tools to search for similar 3D structures. We thus developed an automatic matrix training framework that can generate appropriate substitution matrices for new alphabets when applied in standard 1D sequence alignment methods, e.g. FASTA. Based on the alphabet we constructed, we can transform proteins into 1D structural alphabet representations. To identify protein local structure features, we applied the motif-finding tool MEME [24] to detect the common motifs. We tested two protein families in SCOP, TIM and EGF. The results showed our method successfully recovered their structure domains.

Materials and Methods

The simplest substitution matrix to use is the identity matrix, but it ignores possible acceptable alphabet letter substitutions, which significantly limits its applicability. Some authors applied HMM approach to define the matrix [25], while others adopted a similar approach in the development of BLOSUM matrices [26,27]. Most of these approaches to constructing substitution matrices required the alignments of known proteins [27-29]. As the alignments may be unavailable or even questionable, we took a self-training strategy to build a substitution matrix for our new structural alphabet. This training framework is a flexible and modular design, and it does not rely on any pre-alignment of protein sequences or structures. This matrix training procedure can be applied regardless of how the alphabet is derived. Different training data or alignment tools available can be incorporated in this framework to generate appropriate matrices under various circumstances.

There are three components in the matrix training framework, an alignment tool with a substitution matrix, training data, and a matrix trainer. We used FASTA as the alignment tool, and the non-redundant proteins in SCOP1.69 with sequence similarity less than 40%, excluding the families of size smaller than 5 proteins, as the training dataset. We started by using the identity matrix as the initial substitution matrix where the score is 1 for a match, 0 for a mismatch. Each protein in the training dataset was iteratively used as a query for FASTA to search the rest of the dataset for similar proteins. If a protein returned by FASTA belonged to the same family as the query, we considered the case as a positive hit; otherwise, a negative hit. Those proteins not returned by FASTA but in the same family as the query were considered as misses. For all positive hits and misses, we gathered their alignments with the query produced by FASTA. Based on the alignments, we computed the log-odd ratios defined in the same way as in the BLOSUM matrices [28] to build the *positive matrix*. Similarly, with the alignments of negative hits, we constructed the *negative matrix*. The matrix trainer updated the current substitution matrix $S^{(t)}$ to $S^{(t+1)}$ as the following.

$$S^{(t+1)} = S^{(t)} + M$$

$$M = [W_p \cdot (P - S^{(t)}) - W_n \cdot (N - S^{(t)})] \cdot \tau$$

$$W_p = (|positive\_hits| + |misses|)/|training\_data|$$

$$W_n = |negative\_hits|/|training\_data|$$

where P and N are the positive and the negative matrix respectively, τ is the learning rate (similar to the learning rate in neural networks), and W_p and W_n are the weights. They were defined as the proportion of the total number of positive hits and misses to the training data size and the ratio of the number of negative hits to the training data size, respectively. We repeated the update process to train the substitution matrix until there was no change in the matrix, i.e. the number of both the positive and the negative hits remain constant. The converged matrix was our final substitution matrix which we combined with FASTA as a new alignment tool to demonstrate the applicability of our new alphabet and matrix. We compared our alignment tool with other similar ones on database-scale search tasks. The results were detailed in the next section. The matrix training framework was presented in Figure 1.

Currently, we used the non-redundant proteins in SCOP1.69 with sequence similarity less than 40% for training. We defined the positive hit rate of a query as the ratio of the number of positive hits to the size of the family the query belonged to. As we iterated over each training protein (as a query), we refined the matrix till we could no longer increase the average positive hit rate of all the proteins. One learning example was presented in Figure 2. We tried different learning rates from 0.25 to 1.00. The final average positive hit rates under different learning rates were similar, between 0.9112 and 0.9153. We selected the converged matrix with the maximum positive hit rate when learning rate set 0.50. We named this matrix TRISUM-169 (TRained Iteratively for SUbstitution Matrix-SCOP1.69) as shown in Figure 3.

Experimental Results

Several protein structure search tools based on 1D alignment algorithms have been developed, including SA-Search [25], YAKUSA [30], 3D-BLAST [27], but few were evaluated on the performance of database-scale search. To keep the consistency, we used the same 50 proteins selected from SCOP95-1.69 as used in Yang & Tung’s experiment to compare our alignment tool with 3D-BLAST, PSI-BLAST, YAKUSA MAMMOTH and CE in search time, predictive accuracy and precision. There are some other search tools, e.g. PBE [31], SA-Search [30], Vorolign [32] and so on. Because they either could not be tested on the SCOP database directly (e.g. only PDB available in SA-Search) or the version of their databases provided was older (e.g. ASTRAL in PBE derived from SCOP-1.65, Vorolign server only scans SCOP40-1.69), these tools were not chosen for comparison. We summarized the results in Table 1. It showed that our tool outperformed the other two BLAST-based search tools (i.e. 3D-BLAST and PSI-BLAST) and another structure search tool that also described structures as 1D sequences (i.e. YAKUSA) in predictive accuracy and precision. Compared with the structural alignment tools (i.e. MAMMOTH and CE), our tool obtained a bit worse but comparable accuracy as well as precision. As for search time (using one Intel Pentium 2.8GHz processor and 512Mbytes of memory), Table 1 clearly indicated that our alignment tool was far more efficient than the structural alignment tools, MAMMOTH and CE.

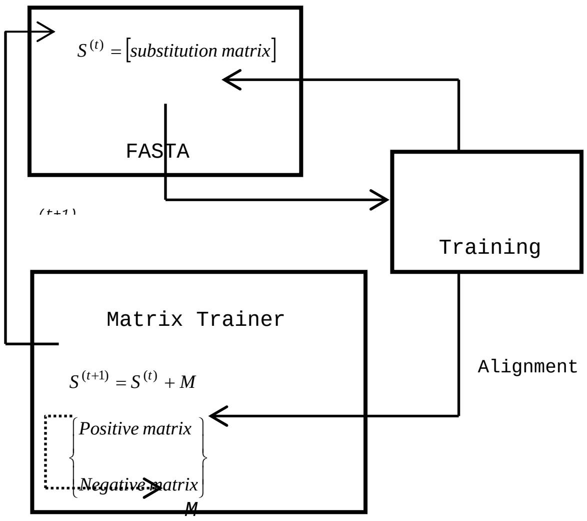

$S^{(t)} = [\textit{substitution matrix}]$
FASTA
Training
Matrix Trainer
Alignment
$S^{(t+1)} = S^{(t)} + M$
Positive matrix
Negative matrix
M

Fig 1. System architecture of the matrix training framework.

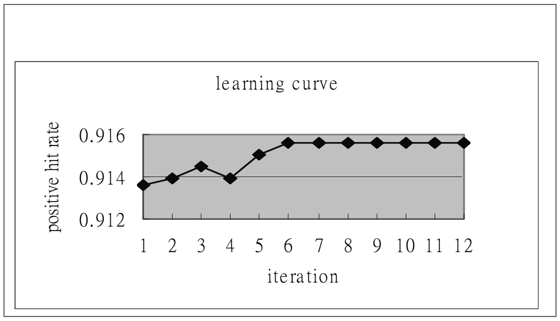
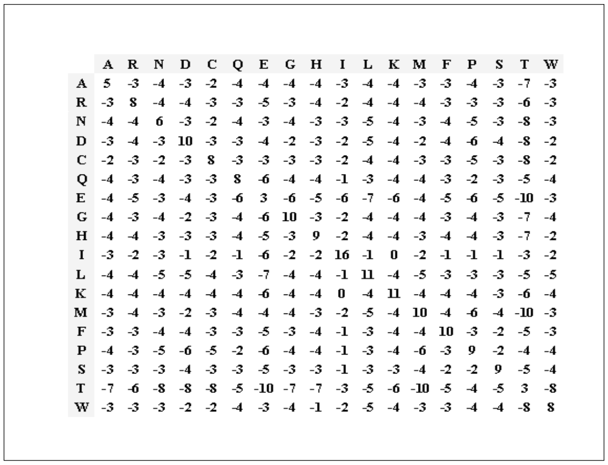


Fig 2. An example of the learning curve of matrix training. The average positive hit rate converged at 0.9153 with the learning rate set 0.5.

To demonstrate the ability of our structural alphabet to describe protein local structure features, we used MEME [24] to detect common motifs in the top 100 hits found by our alignment tool. These motifs could be well mapped to the eight β/α barrel strands of TIM barrel domains. Figure 4(a) showed the structure of archaeon pyrococcus woesei (PDB 1hg3a). In Figure 4(b), we highlighted the identified motif in PDB 1hg3a, and Figure 4(c) illustrated the motif structure. The structural alphabet letter sequence of this motif and the corresponding amino acids were shown in Figure 4(d). In addition to TIM barrel structures, we also used the EGF/EGF-like domain as another study case. Epidermal growth factor (EGF) domains are extracellular protein modules typically described by 30-40 amino acids primarily stabilized by three disulfide bonds. Compared with TIM barrel structures, EGF are much smaller domains. We used it to evaluate how well a structural alphabet could define the 3D structures of small proteins. Many proteins contain the regions of homology to EGF, and the cysteine residues at similar positions. The homologies and available functional data suggest that these domains share some common functional features. If we number the cysteine residues as Cys1 to Cys6, where Cys1 is the closest to the N-terminus, the regularity of cysteine spacing defines three regions, A, B and C. Based on the conservation in sequence and length of these regions, the homologies have been classified into three different categories [33]. We described the 227 proteins in the EGF-type module family of SCOP 1.69 in our alphabet, Yang & Tung's [27] and de Brevern *et al.*'s [15,26,31], respectively. We then used MEME to identify the common motifs corresponding to the sub-domains, A, B and C. According to InterPro [34], 24 of these proteins were exclusively of *EGF Type-1*, 74 were of *EGF-like Type-2*, and 117 belonged to *EGF-like Type-3* only. We classified the remaining 12 proteins as *Others*.

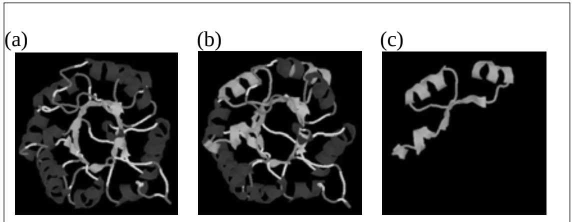


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W
A	5	-3	-4	-3	-2	-4	-4	-4	-4	-3	-4	-4	-3	-3	-4	-3	-7	-3
R	-3	8	-4	-4	-3	-3	-5	-3	-4	-2	-4	-4	-4	-3	-3	-3	-6	-3
N	-4	-4	6	-3	-2	-4	-3	-4	-3	-3	-5	-4	-3	-4	-5	-3	-8	-3
D	-3	-4	-3	10	-3	-3	-4	-2	-3	-2	-5	-4	-2	-4	-6	-4	-8	-2
C	-2	-3	-2	-3	8	-3	-3	-3	-3	-2	-4	-4	-3	-3	-5	-3	-8	-2
Q	-4	-3	-4	-3	-3	8	-6	-4	-4	-1	-3	-4	-4	-3	-2	-3	-5	-4
E	-4	-5	-3	-4	-3	-6	3	-6	-5	-6	-7	-6	-4	-5	-6	-5	-10	-3
G	-4	-3	-4	-2	-3	-4	-6	10	-3	-2	-4	-4	-4	-3	-4	-3	-7	-4
H	-4	-4	-3	-3	-3	-4	-5	-3	9	-2	-4	-4	-3	-4	-4	-3	-7	-2
I	-3	-2	-3	-1	-2	-1	-6	-2	-2	16	-1	0	-2	-1	-1	-1	-3	-2
L	-4	-4	-5	-5	-4	-3	-7	-4	-4	-1	11	-4	-5	-3	-3	-3	-5	-5
K	-4	-4	-4	-4	-4	-4	-6	-4	-4	0	-4	11	-4	-4	-4	-3	-6	-4
M	-3	-4	-3	-2	-3	-4	-4	-4	-3	-2	-5	-4	10	-4	-6	-4	-10	-3
F	-3	-3	-4	-4	-3	-3	-5	-3	-4	-1	-3	-4	-4	10	-3	-2	-5	-3
P	-4	-3	-5	-6	-5	-2	-6	-4	-4	-1	-3	-4	-6	-3	9	-2	-4	-4
S	-3	-3	-3	-4	-3	-3	-5	-3	-3	-1	-3	-3	-4	-2	-2	9	-5	-4
T	-7	-6	-8	-8	-8	-5	-10	-7	-7	-3	-5	-6	-10	-5	-4	-5	3	-8
W	-3	-3	-3	-2	-2	-4	-3	-4	-1	-2	-5	-4	-3	-3	-4	-4	-8	8

Fig 3. Substitution matrix TRISUM-169.

Despite that the sub-domains are less conserved in EGF-like Type-3, sub-domain A is typically composed of five to six residues in

Type-1 and 2, sub-domain B usually contains 10-11 residues in Type-1, but consistently three residues shorter than in Type-1, sub-domain C is conserved in length with four or five specific residues in Type-1 and 2 [33]. We used 8, 10 and 15 respectively as the motif width and ran MEME to find motifs. A motif found was considered as corresponding to a sub-domain correctly if more than half of the residues in the sub-domain were included in the motif. If any single motif of width 8, 10 or 15 alphabet letters correctly corresponded to a sub-domain, we claimed this sub-domain was recovered successfully (i.e. a hit). We summarized the results of the motifs found in Table 2. It showed that with our structural alphabet MEME was able to identify more EGF sub-domains than using Yang & Tung's or de Brevern *et al.*'s alphabets.



(d)

```
      1       10        20        30        40        50
       |--------|---------|---------|---------|---------|
(AA) AKLKEPIIAINFKTYIEATGKRALEIAKAAEKVYKETGVTIVVAPQLVDL
(SA) NNEACWNEEEMRQSFRQSFRQTTTTTTTTTTTTTSLKGHNEEEEEARQPT
     51       10        20        30        40       100
      |--------|---------|---------|---------|---------|
(AA) RMIAESVEIPVFAQHIDPIKPGSHTGHVLPEAVKEAGAVGTLLNHSENRM
(SA) TTTTPSFCWNEEACWNEEEMADWLADHNARQTTTLKGHRCWNADARQSNE
     101       10        20        30        40      150
      |--------|---------|---------|---------|---------|
(AA) ILADLEAAIRRAEEVGLMTMVCSNNPAVSAAVAALNPDYVAVEPPELIGT
(SA) ARQTTTTTTTTTTLKGHNEEEEACARQTTTTTTLKGFCWNEEEARQPPLK
     151       10        20        30        40      200
      |--------|---------|---------|---------|---------|
(AA) GIPVSKAKPEVITNTVELVKKVNPEVKVLCGAGISTGEDVKKAIELGTVG
(SA) GHFRQLKGQTTTTTTTTTTTTTSPPCNNEEMMDDHCARQTTTTTTLKGLR
     201       10        20        30        40      250
      |--------|---------|---------|---------|---------|
(AA) VLLASGVTKAKDPEKAIWDLVSGI
(SA) DWNARQTTTSFCQPQTTTTTTTPP
```

Fig. 4. Common motif found by MEME in PDB 1hg3a. (a) TIM barrel structure of PDB 1hg3a (b) motif highlighted in green (c) motif structure (d) PDB 1hg3a described in amino acids (AA) and structural alphabet (SA), respectively, where motif underlined. (Note. Images are shown in grey scale.)

4 Discussion

The protein structure data we used to build the alphabet were from the non-redundant PDB database instead of some specialized databases, e.g. Pair Database [27] and PDB-SELECT [29], with the aim to ensure the generality of our alphabet. We also proposed an automatic matrix training framework to construct an appropriate substitution matrix for the alphabet. This training strategy did not need any information of known alignments that most previous works required. Using different training data and update rules, the self-training methodology can be applied to various alphabets.

To demonstrate the performance of our alignment tool, we systematically compared it with other search tools. The results showed that our new tool was very competitive in predictive accuracy and alignment efficiency for database-scale search. We further evaluated the potential of using motif-finding tools, e.g. MEME, to detect structure domains/sub-domains represented in our structural alphabet. Two examples of different protein classes, TIM in α/β and EGF in small proteins, have been tested. The results indicated that the identified motifs mapped well to the known structure sub-domains.

We can extend the work in several directions. First, we can use a more complete datasets for substitution matrix training to increase sensitivity and selectivity in database search. Second, besides FASTA, we can combine other alignment tools with our substitution matrix, and evaluate the performance of different combinations. Third, currently we use MEME to detect motifs, and we have demonstrated it is able to recover some structure sub-domains described in our structural alphabet. MEME was originally designed to find motifs in amino acid and nucleic acid sequences. To increase the performance in structural motif detection, we can either modify MEME or develop a new motif-finding tool specifically for our structural alphabet. Finally, several structural alphabets have been developed based on different protein structural characteristics. It is worthwhile to conduct a thorough comparative study and evaluate the feasibility of combining different alphabets. The combination of structural alphabets that complement each other will increase their overall applicability and characterize 3D protein structures more completely.

5 References

[1] A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures", *J. Mol. Biol.*, 1995, 536-540.

[2] R. Unger, D. Harel, S. Wherland and J.L. Sussman "A 3D building blocks approach to analyzing and predicting structure of proteins", *Proteins*, 1989, 355-373.

[3] M. Dudev and C. Lim "Discovering structural motifs using a structural alphabet: Applications to magnesium-binding sites", *BMC Bioinformtics*, 2007, 106.

[4] R. Aurora, R. Srinivasan and G.D. Rose "Rules for alpha-helix termination by glycine", *Science*, 1994, 1126-1130.

[5] R. Unger and J.L. Sussman "The importance of short structural motifs in protein structure analysis", *J. Comput. Aided Mol. Des.*, 1993, 457-472.

[6] Z.Y. Zhu and T.L. Blundell "The use of amino acid patterns of classified helices and strands in secondary structure prediction", *J. Mol. Biol.*, 1996, 261-276.

[7] K.F. Han and D. Baker "Recurring local sequence motifs in proteins", *J. Mol. Biol.*, 1995, 176-187.

[8] K.T. Simons, C. Kooperberg, E. Huang and D. Baker "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions", *J Mol Biol.*, 1997, 209－225.

[9] J. Garnier, D. Osguthorpe and B. Bobson "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular protein", *Journal of Molecular Biology*, vol. 120, 1978, pp. 97-120.

[10] B. Rost and C. Snader, "Prediction of protein secondary structure at better than 70% accuracy", *Journal of Molecular Biology*, vol. 232, 1993, pp. 584-599.

[11] A. Salamov and V. Solovyev, "Protein secondary structure prediction using local alignments", *Journal of Molecular Biology*, vol. 268, 1997, pp. 31-36.

[12] T.N. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G.P. Gippert and O. Lund, "Prediction of protein secondary structure at 80% accuracy", *Proteins*, vol. 41, 2000, pp. 17-20.

[13] B. Rost, "Review: Protein secondary structure prediction continues to rise," *Journal of Structural Biology*, vol. 134, 2001, pp. 204-218.

[14] A.G. de Brevern and S.A. Hazout, "Hybrid Protein Model(HPM): a method to compact protein 3D-structure information and physicochemical properties", *IEEE Comp. Soc*. S1, 2000, pp. 49-54.

[15] A.G. de Brevern, H. Valadie, S.A. Hazout and C. Etchebest, "Extension of a local backbone description using a structural alphabet: A new approach to the sequence-structure relationship," *Protein Science*, vol. 11, 2002, pp. 2871-2886.

[16] R. Unger, D. Harel, S. Wherland and J.L. Sussman, "A 3D building blocks approach to analyzing and predicting structure of proteins", *Proteins*, vol. 5, 1989, pp. 355-373.

[17] J. Schuchhardt, G. Schneider, J. Reichelt, D. Schomburg and P. Wrede, "Local structural motifs of protein backbones are classified by self-organizing neural networks", *Protein Engineering*, vol. 9, 1996, pp. 833-842.

[18] M.J. Rooman, J. Rodriguez and S.J. Wodak, "Automatic definition of recurrent local structure motifs in proteins", *Journal of Molecular Biology*, vol. 213, 1990, pp. 327-336.

[19] J.S. Fetrow, M.J. Palumbo and G. Berg, "Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme", *Proteins*, vol. 27, 1997, pp. 249-271.

[20] C. Bystroff and D. Baker, "Prediction of local structure in proteins using a library of sequence-structure motif", *Journal of Molecular Biology*, vol. 281, 1998, pp. 565-577.

[21] A.C. Camproux, R. Gautier and P. Tuffery, "A hidden Markov model derived structural alphabet for proteins", *Journal of Molecular Biology*, doi: 10.1016/j.jmb.2004.04.005.

[22] S. Ku and Y. Hu "A Multi-strategy Approach to Protein Structural Alphabet Design", *Biocomp* 2006.

[23] W.R. Pearson "Flexible sequence similarity searching with the FASTA3 program package", *Methods Mol. Biol.*, 2000, 185-219.

[24] T.L. Bailey and C. Elkan "Unsupervised learning of multiple motifs in biopolymers using EM", *Machine Learning*, 1995, 51-80.

[25] F. Guyon, A.C. Camproux, J. Hochez and P. Tuffery "SA-Search: a web tool for protein structure mining based a structural alphabet", *Nucleic Acids Res.*, 2004, W545–W548.

[26] M. Tyagi, V.S. Gowri, N. Srinivasan, A.G. de Brevern and B. Offmann “A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications”, *Proteins: Structure, Function and Bioinformatics*, 2006, 32-39.

[27] J.M. Yang and C.H. Tung “Protein structure databases search and evolutionary classification”, *Nucleic Acids Research*, 2006, 3646-3659.

[28] S. Henikoff and J.G. Henikoff “Amino acid substitution matrices from protein blocks”, *PNAS*, 1992, 10915-10919.

[29] W.M. Zheng and X. Liu “A protein structural alphabet and its substitution matrix CLESUM”, *LNCS*, 2005, 59-67.

[30] M. Carpentier, S. Brouillet and J. Pothier “YAKUSA: a fast structural database scanning method”, *Proteins: Structure, Function and Genetics*, 2005, 137-151.

[31] M. Tyagi, P. Sharma, C.S. Swamy, F. Cadet, N. Srinivasan, A.G. de Brevern and B. Offmann “Protein Block Expert (PBE): a web-based protein structure analysis server using structural alphabet”, *Nucleic Acids Research*, 2006, W119-123.

[32] F. Birzele, J.E. Gewehr, G. Csaba and R. Zimmer “Vorolign- fast structural alignment using Voronoi contacts”, *Bioinformatics*, 2007, e205-211.

[33] E. Appella, I.T. Weber and F. Blasi “Structure and function of epidermal growth factor-like regions in proteins”, *FEBS Letters*, 1988, 1-4.

[34] N.J. Mulder *et al.* “New developments in the InterPro database”, *Nucleic Acids Research*, 2007, D224-228.

Table 1. Comparison between our alignment tool, 3D-BLAST, PSI-BLAST, YAKUSA, MAMMOTH and CE on 50 proteins selected fromSCOP95-1.69.

Search tool	Average time required for a query (sec)	Relative to SA-FAST	Accuracy (%)	Average precision (%)
Our Tool	1.15	1.00	96	90.80
3D-BLAST	1.30	1.13	94	85.20
PSI-BLAST	0.48	0.42	84	68.16
YAKUSA	8.88	7.72	90	74.86
MAMMOTH	1834.18	1594.94	100	94.01
CE	22053.32	19176.80	98	90.78

Table 2. Comparison between our structural alphabet, Yang & Tung's and de Brevern *et al.*'s in describing motifs found by MEME within EGF family.

(a) Number of motifs found by MEME, using different structural alphabets to describe EGF (EGF-like) proteins

		Our SA						**Yang & Tung's**						**de Brevern *et al.*'s**					
Sub-domain Type		**A**		**B**		**C**		**A**		**B**		**C**		**A**		**B**		**C**	
EGF proteins	**No.**[a]	**Hits**[b]	**Cov**[c]	**Hits**	**Cov**	**Hits**	**Cov**	**Hits**	**Cov**	**Hits**	**Cov**	**Hits**	**Cov**	**Hits**	**Cov**	**Hits**	**Cov**	**Hits**	**Cov**
Type 1	24	23	95.8	22	91.7	23	95.8	11	45.8	21	87.5	19	79.2	18	75.0	14	58.3	18	75.0
Type 2	74	73	98.6	71	95.9	74	100.0	62	83.8	73	98.6	60	81.1	68	91.9	62	83.8	70	94.6
Type 3	117	116	99.1	106	90.6	61	52.1	54	46.2	102	87.2	25	21.4	109	93.2	112	95.7	48	41.0
Others	12	12	100.0	11	91.7	11	91.7	9	75.0	11	91.7	9	75.0	12	100.0	11	91.7	9	75.0
All	227	224	98.6	210	92.5	169	74.4	136	59.9	207	91.2	113	49.8	207	91.2	199	87.7	145	63.9

[a] The number of EGF proteins of a specific type, [b] We called it a hit for a sub-domain when more than half of the sub-domain residues were contained in a motif. We presented the count of hits of different types, [c] Cov(Coverage) was defined as the ratio of the count of hits to the number of EGF proteins, e.g., if No.=24 and Hits=22, then Cov=22/24=91.7%.

(b) Statistics of EGF (EGF-like) proteins whose sub-domains detected by MEME

	Structural Alphabet					
EGF proteins	**Our SA**		**Yang & Tung's**		**de Brevern *et al.'s***	
	Count	**Percentage**	**Count**	**Percentage**	**Count**	**Percentage**
Found 3[a]	151	66.52	79	34.80	104	45.81
Found 2[b]	74	32.60	78	34.36	116	51.10
Found 1[c]	2	0.88	63	27.75	7	3.08
Found 0[d]	0	0.00	7	3.08	0	0.00
Total	227	100.00	227	100.00	227	100.00

[a] EGF (EGF-like) proteins in which all three sub-domains (A, B and C) were found by MEME, [b] EGF (EGF-like) proteins in which two out of three sub-domains were found by MEME, [c] EGF (EGF-like) proteins in which only one sub-domain was found by MEME, [d] EGF (EGF-like) proteins in which MEME failed to identify any sub-domain.