Song Huang **CBB752**

Sequencing and comparison of yeast species to identify genes and regulatory elements

Introduction:

With the availability of huge amount of biological sequencing data, how to identify the functional elements encoded in a genome is one of the principal challenges in modern biology (1). In this paper, Kellis et al. addressed this question and presented a comparative genomics approach for systematic gene and regulatory elements identification. They compared the high-quality draft genome sequences of three yeast species *S.paradoxus, S. mikatae and S.bayanus* to their close relative, *S. cerevisiae*(1).

They first aligned the genomes and characterized their evolution, defining the regions and mechanisms of change (1). Then they presented methods for the identification of protein-coding genes based on their patterns of nucleotide conservation across related species (1, 2). They also presented the novel methods for the systematic *de novo* identification of regulatory motifs. The methods do not rely on previous knowledge of gene function and in that way differ from the existing computational methods on motif discovery (1, 2).

Genome alignment

Before the gene and motif finding analysis, they need to identify which genomes to use and how to compare these genomes. The optimal choice that they considered includes: First, the branch length *t* between species should be short enough to permit orthologous sequences to be readily aligned. Second, the total branch length of the tree should be large enough that non-functional sites will have undergone substantially more

drift than functional sites, there by providing an adequate degree of signal-to-noise enrichment. Third, the species should represent as narrow a taxon as possible (1). Based on these considerations, the yeast species they studied have t = 0.23-0.55, total branch length of the phylogenic tree is 0.83, and the identified elements shared across *Saccharomyces sensu stricto*.

To identify orthologous regions, they used a multiple sequence alignment algorithm (2) to build ORF synteny blocks and get one-to-one orthologous matches in each species (2). They analyzed all of the possible ORFs including 6,235 ORFs in the current SGD annotation for *S. cerevisiae*. Only 211 ORFs shows ambiguous correspondences and can be explained by gene-family expansion/contraction using the clustering data.

Identification of genes

Based on Kellis et al. (3), they used a classification-based approach for systematic gene identification using genome sequence comparison. The principle is simple: test the ORFs seen in one species by observing whether the orthologous sequence in related species also encodes an ORF. True protein-coding ORFs will typically be under strong selective pressure to preserve the open reading frame, whereas spurious ORFs will accumulate frame shift and stop codons (1).

They first build the *synteny blocks* and construct global nucleotide alignments. With the availability of orthologous alignments for both protein-coding and non-coding regions, they built a classifier between the two types of region and applied the same test (3). They developed reading frame conservation (RFC) test to classify each ORF in

S.cerevisiae as biologically meaningful or meaningless, on the basis of the proportion of the ORF over which reading frame is locally conserved in each of the other three species.

The authors identified 6,275 ORFs (6,062 in SGD) that could theoretically encode proteins of larger than 100 amino acids and do not overlap a longer ORF by more than half of their length. They applied the RFC test to all of the 6,602 ORF in SGD. With the removal of some ambiguity and short ORFs, they proposed a revised yeast gene catalogue consisting of 5,538 ORFs encoding protein at least 100 amino acid. About 98% of the genes identified agree with annotation and only one true gene fail on RCF test showed the high sensitivity and specificity of their method.

Genome-wide identification of regulatory elements

They developed statistical methods for the systematic de-novo identification of regulatory motifs. Without previous biological knowledge, we discovered virtually all previously known DNA regulatory motifs as well as several noteworthy novel motifs. With the additional use of gene ontology information, expression clusters and transcription factor binding profiles, they assigned candidate functions to the novel motifs discovered (4).

They defined motif conservation score (MCS) on the basis of the conservation rate of the motif in intergenic regions: the MSC is measured in standard deviations above the rate for comparable control motifs (1). To discover motif, they first identify conserved 'mini-motifs' with pattern $XYZn_{(0-20)}UVW$. Conserved mini-motifs are defined according to three conservation criteria: (1) intergenic conservation; (2) intergenic-genic conservation; (3) upstream-downstream conservation. Then they use the

mini-motif identified to construct full motifs. As ‘motif-cores’ (2), those mini-motifs are first extended by searching for nearby sequence positions showing significant correlation with a mini-motif. The extended motifs are then clustered, merging those with substantially overlapping sequences and those that tend to occur in the same intergenic regions. Finally, a full motif is created by deriving a consensus sequence (1).

About 2,400 mini-motifs show high scores by one or more of the conservation criteria and give rise to a list of 72 full motifs having MCS>= 4. These 72 discovered motifs show strong overlap with 28 of the 33 known motifs having MCS>=4 and 8 of the 22 known motifs having MCS<4. Comparative genomic analysis thus automatically discovered 36 known motifs and 42(?) new motifs.

They have used the extensive experimental knowledge in yeast to validate our results, thus confirming that the methods presented here are applicable to other species.

Discussion:

The goal of this paper is to develop and apply general approaches for systematic analysis of protein-coding and regulatory elements within any genome by means of whole-genome comparisons with several related species (1). Their analysis provides a *de novo* way to view the dynamic nature of cell (2) without knowing too much previous gene function knowledge. The high discovery sensitivity and specificity of gene and motif finding results give the authors high confidence to claim that comparative analysis with closely related species can be invaluable in understanding a genome (2).

This is the first to use sequence comparative methods for gene finding and motif finding at genome level. One of the merits of this analysis is to align different genomes at

nucleotide level. Then the RFC test provides a classifier between coding and non-coding regions that does not rely on start, stop or splicing signals, nor does it rely on the conservation of protein sequence (3). In this way, the analysis results only depend on the sequences themselves.

. The second good thing about is that the genome comparison method will be effected by experimental noise less than other traditional methods. For example, with certain experimental conditions, some gene may never be found expressed, or due to cross hybridization, the real transcriptional signal cannot be distinguished from background noise level. Whereas, the genome comparative methods presented in this paper will only consider the annotated sequences only and using statistical models to test the discovery significance.

One concern I used to have when I first read the abstract of this paper is how to deal with those rapid evolutionary genes. The authors addressed this question in the paper and explained some results in this case.

Sequence alignment is still one of the most basic and powerful tool in bioinformatics field. With the solid biological knowledge on gene regulatory and good understanding on genome sequence analysis, the authors provide a nice way to treat the genome sequence as pure data without considering previous biological knowledge other than validating their results. This is a paper every bioinformatics should read and grasp the idea how to use bioinformatics as a tool to facilitate biology research.

Extension: Discover the relationship between regulatory elements and genes in yeast

The goal is to use Quantitative Trait Loci (QTL) linkage analysis method and expression data to reveal the regulatory relationship between motifs and genes in yeast.

In the paper discussed above, the authors have identified genes and motifs genome-wide and compared with the existed annotations databases. With some additional data, we can carry some linkage analysis to reveal the relationship between the causality location and trait location; therefore, we can possibly build the regulatory network for yeast (5, 6).

Dataset and software needed:

Current available datasets: 1) the *S. cerevisiae* genes and motifs identified by Kellis et al. (2003); 2) the revised genome annotation corresponding to their results; 3) not revised yeast genome data as reference.

Additional datasets needed: 4) *S. cerevisiae* expression data; 5) *S. cerevisiae* genotype data;

The dataset 4) can be obtained from NCBI's database gene express OmniBus and dataset 5) can be downloaded from Dr. Kruglyak's website (Brem et al. 2002). The expression dataset contains the expression levels of all of the annotated genes. The genotype dataset contains about 3000 genetic (SNP) markers. Totally there are 40 haploids in the dataset and they are the second backcross offspring from laboratory (BY) and wild (RM) strains.

Software: some QTL analysis software package, such Rqtl, mapmaker, etc.

Mechanism to detect the causal elements' location and procedures

The traditional QTL approach is used to detect the gene causing a certain disease (phenotype). For animal study, researches usually take advantage of inbred line for the simplicity. Suppose we have yeast strain A & B with very distinctive phenotypes and both A & B are inbred strain with genotype AA and BB, respectively. After mating A and B, the F1 offspring have genotype AB only. Then backcrossing F1 with one of its parents, say, A, we can obtain B2 generation with genotype AB or AA at different locations due to chromosome crossover. By studying the relationship between the segregation of chromosome and the corresponding phenotype level of each B2 haploid, we can detect the genetic location affecting the phenotype.

Similar to the traditional QTL approach, we can use the gene expression level as a quantitative trait, which may be effected by more than one causal element (motif, gene, etc). We can map the possible causal loci by scanning along the genome and testing the correlation between the expression levels of every gene with the genotype of the current marker. If the gene and marker currently tested show significant correlation, we can claim that there is a causal element around the marker and effecting the expression of the gene tested.

About the statistical testing, we can use the simplest ANOVA test first to get preliminary results. We can divide all of the haploids into two groups according to their genotypes (AB or AA) for the current marker tested. Then we build ANOVA table for their expression values and test whether the two groups have different mean value. Significant difference of two group means indicates causal locus is found. We can scan

the whole genome to find which gene the current marker might affect. Then we go to next marker to do another round of genome scan.

Difficulties:

The primary difficulty of this study will be the resolution of the marker. The marker already genotyped maybe not located close to any annotated motif. Since we need to detect the relationship between the gene expression and founded motifs, the genetic markers used should close to the regulatory element to give high significance of testing. Since the genotyping for the current dataset is done, we can't improve this other than finding more complete dataset.

False positive and false negative: For the motifs that far way from all of the markers, we might get false negative result. On the other hand, since motifs are not the only cause of expression level, we may found 'new motif' showing significance. For example, transcription factor gene coding region might show strong relationship with many other genes. These regulatory relate products will give false positive to our motif testing.

Other approach might be used to validate the results:

Yeast Chip-chip data might also be used to validate the positions of regulatory elements. One of my concerns is that the low resolution chip-chip method of might only be able to provide weak support the motif localization. Also, we can't build the relationship between the regulatory positions with target genes (4).

References:

1. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241-54 (2003).

2. Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J Comput Biol* 11, 319-55 (2004).

3. Manolis Kellis, Gene finding using multiple related species: a classification approach. Special review for the Encyclopedia of Genomics, Proteomics, and Bioinformatics. John Wiley and Sons, editors. (in press)

4. X. Shirley Liu, Douglas L. Brutlag, & Jun S. Liu. Analgorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments, *Nature*, 2002

5. Rachel B. Brem, John D. Storey, Jacqueline Whittle and Leonid Kruglyak. Genetic interactions between polymorphisms that affect gene expression in yeast, *Nature*, 2005

6. Nan Bing, Ina Hoeschele, Genetical Genomics Analysis of a Yeast Segregant Population for Transcription Network Inference, *Genetics*, 2005