YEAST COMPARATIVE GENOMICS

1. Introduction

A fundamental aspect of genome analysis is the discrimination of functional elements in the genome—e.g. genes, introns, exons, regulatory elements, structural elements—from nonfunctional elements. Before the genomic revolution produced abundant DNA data on a multitude of species, methods for functional element discovery were of two types. 1) The first use as input both the DNA sequences being analyzed and expression data—e.g. cDNA sequences, data from DNA microarray experiments. 2) The second, called *de novo* methods, use only DNA sequences as input. *De novo* methods are preferred, since reliance on prior knowledge is both costly and precludes easy generalizability. Both methods have enjoyed moderate success, but neither has yet produced satisfactorily complete and reliable genome analyses.[1]

This paper introduces a novel *de novo* approach to functional element discovery, comparative genome analysis (CGA). The central premise of CGA is that conservation across closely related species can serve as a signal for functional elements. That is, since an absence of strong selective pressure on nonfunctional sequences makes them more prone to drift than functional sequences, functional sequences should be recognizable based on their higher degree of conservation. The advantage of the CGA approach is that it utilizes a wealth of information provided by evolution to detect biologically meaningful patterns but, like traditional *de novo* approaches, does so without a costly reliance on prior knowledge.

2.1 Purpose

This is a method development paper. The paper's primary purpose is to develop approaches for systematic functional element discovery, via genome comparisons of closely related species. Its secondary purpose is to use these methods to help interpret genomes.

2.2 Dataset

The CGA methods developed here were tested on the well studied eukaryote, *S. cerevisiae*. Sequences of three related species from the *Saccharomyces* genus—*S. paradoxus*, *S. mikatae*, *S. bayanus*—were used for cross genome comparisons. These three species were chosen for comparison because their genomes are similar enough to *S. cerevisiae* to allow for good ortholog alignment, at both the genome level and nucleotide level. And equally important was that they are also diverged enough from *S. cerevisiae* to allow for recognition of functional elements based on degree of conservation.

The genome sequence for diploid *S. cerevisiae* was a finished version, obtained in May 2002 from the Saccharomyces Genome Database (http://www-genome.stanford.edu/Saccharomyces/). The genome sequences for diploid *S. paradoxus*, *S. mikatae*, and *S. bayanus* were high quality draft versions obtained from E. Louis at the University of Leicester, probably around May 2002 as well.

2.3 Methods

Preliminaries

Critical to all methods of the CGA approach is ortholog alignment. So first, the *S. cerevisiae* genome and other species' genomes were aligned according to Blast hits to determine general synteny blocks and one-to-one orthologous ORF maps. After this preliminary work, the authors used the alignment to devise information extracting CGA tests. The most substantial components of this paper use CGA are now considered.

(i) Gene Identification: Reading Frame Conservation Test

Since it is easy enough to search the genome for continuous sequences enclosed by a start and stop codon, the primary challenge in *de novo* gene identification is discriminating between ORFs that correspond to actual genes (true ORFs) and those that do not (false ORFs). The CGA method attempts to resolve the ORF ambiguity in one species by testing the orthologous sequences of other species for ORF conservation. Functional ORFs will have a strong selective pressure to conserve the ORF. Nonfunctional (spurious) ORFs will have marginal selective pressure and thus will be less likely to conserve the ORF, by developing frameshifts and accumulating additional stop codons. The Reading Frame Conservation (RFC) test characterizes this intuition and gives an approximate conservation score; high score implies the *S. cerevisiae* ORF is legitimate; low implies spurious. See the additional figure in Section 3 for elucidation of the RFC test procedure and a hypothetical example.

(ii) Regulatory Element Identification: Motif Conservation Score

Due to the shortness of regulatory sequences and the deficiency of established rules about them, regulatory element identification poses even more challenges than does gene identification. Moreover, in any one genome, many motifs occur frequently by chance. The CGA approach simplifies motif discovery and reduces single genome noise by locating small, frequent regions of atypically high sequence conservation between species. The Motif Conservation Score characterizes this intuition by assigning a score to each proposed motif m based on the conservation of m in intergenic regions.

These proposed motifs were discovered by first constructing all permutations of miniature motifs, searching for conserved mini-motifs, and extended them with conserved bases to generate full motifs. Finally, an MCS value was assigned to the full motif to demonstrate the likelihood of its being a legitimate functional element.

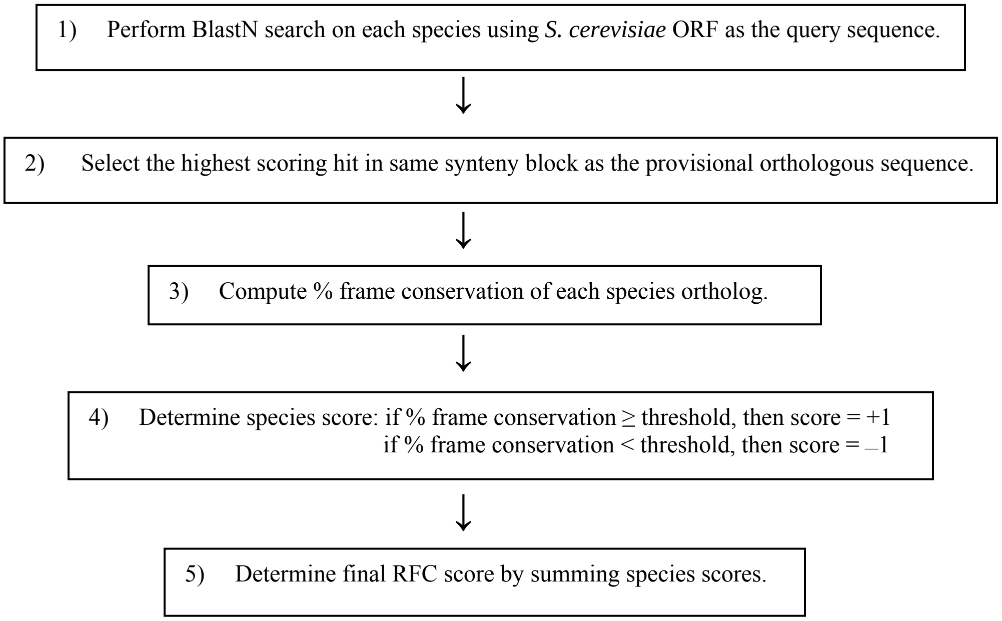
2.4 Results and Conclusions

A primary result of the paper was a significant revision of the yeast gene catalogue, which affected 15% of all genes and added 17 regulatory motifs. The paper covered vast ground, and also addressed some ways CGA can be used to interpret genomes. Effective methods were also developed to define gene structure, identify regions of fast and slow evolutionary change, infer functions of motifs based on gene categories, and demonstrate combinatorial control of gene regulation via motif 'interactions.'

3. Additional Figure: Elucidation of RFC Test Procedure

a)

General Scheme of RFC Test Procedure



1) Perform BlastN search on each species using *S. cerevisiae* ORF as the query sequence.

2) Select the highest scoring hit in same synteny block as the provisional orthologous sequence.

3) Compute % frame conservation of each species ortholog.

4) Determine species score: if % frame conservation ≥ threshold, then score = +1
if % frame conservation < threshold, then score = –1

5) Determine final RFC score by summing species scores.

b)

Output of RFC Test Procedure for Hypothetical *S. cerevisiae* ORF YBR032K

Ortholog Sequence	*S. paradoxus* unnamed	*S. mikatae* unnamed	*S. bayanus* unnamed
% Reading Frame Conservation	86%	54%	63%
Species Threshold	80%	75%	70%
Species Score	+1	-1	-1

Final RFC Score = -1

Low RFC score implies YBRO32K is a specious ORF

4. Additional Comments

The insight of this paper was far reaching. Aside from just identifying genes and regulatory elements, comparative genomics methods are also probably the best approach to understanding how gene and regulatory element position in the genome affect their contribution to cellular life. And since position and synteny are of crucial importance in dynamic gene expression, CGA approaches on this topic might open doors in areas where a better understanding dynamic gene expression is necessary for future development, like genetic therapy. Also, as will be discussed shortly, DNA microarray experimentation seems as though it could contribute greatly to the efficacy of the CGA methods found in this paper.

5. Proposed Extension: CGA and Gene Expression Profiling to Discover Regulatory Motif Networks Used for Combinatorial Control in Differential Gene Expression

Introduction: A key question in modern biology is how cells achieve differential gene expression. This question includes topics like how cellular differentiation occurs and how already differentiated cells achieve specific gene expression responses to a vast array of external stimuli. At the genetic level, possible modes of regulation include gene position, gene frequency, combinatorial control of transcription factors, and combinatorial control of the regulatory motifs that transcription factors bind. The topic of motif combinatorial control is especially interesting and complex because cells tend to have such a limited number of motifs—e.g., *S. cerevisiae* has only 72 known regulatory motifs according to the CGA paper. Thus, if cells use this method of control, regulatory motifs are likely to form intricate networks with each other, via cooperative and antagonistic 'interactions.'

Presently, the general approach to uncover the motif network is to search for motifs whose presence in particular intergenic regions correlate. But the motif search space in genomes is vast, complex, and subtle. As discovered in the CGA paper, the tremendous amount of biological noise prevents such correlations from being significant in single genome studies. Comparative genomics alleviates the noise problem somewhat by using conservation as a criterion for functional elements. But the subtlety of the genome's operations makes it likely that much more of the motif network has yet to be discovered. One way to further uncover the motif network is to combine the CGA approach with the enormous experimental power of DNA microarrays.

Employing gene expression profiles could be especially fruitful since they contain within them the information that motif networks contribute to differential gene expression. A loose analogy of motif network discovery to protein structure prediction comes to mind: motif containing intergenic sequences are to experimentally determined gene expression profiles as amino acid sequences are to experimentally determined structures. The experimental data shows the final results of member interactions and can give dramatic clues as to where to look for those interactions.

Accordingly, the main line of thought of the following proposed study is to use empirical studies to reduce the search space for motif 'interactions' and to use cross species comparisons to reduce biological noise; both of these results should increase the sensitivity of methods which find motif correlations and 'interactions.' The basic strategy of the combined approach is: perform a cross-species comparison of gene expression profiles induced by the same stimulus to determine which genes are expressed in different ways. The intergenic space around these differentially expressed genes will be analyzed for differences in motif population that show signs of motif 'interactions.'

Purpose: The purpose of this study is to develop methods for uncovering the regulatory motif network, using the ideas and methods of comparative genomics and gene expression profiling.

Dataset: Much work in both comparative genomics and gene expression profiling has already been performed on *S. cerevisiae*, thus it will be the object of study.[2,3] The finished genome sequence of *S. cerevisiae* and the high quality draft sequences of *S. paradoxus*, *S. mikatae*, and *S. bayanus* will be used from the *Saccharomyces* Genome Database (http://genome-www.stanford.edu/Saccharomyces/).

Methods: Three methods will be considered, in increasing order of complexity. Also, the concept of gene categories, that is clusters of genes that are often expressed together or serve similar functions, is crucial for all of the methods. In this study, either traditional functional categories or simultaneous expression categories, like those developed by Hughes et al., can be used.[3]

(1) *Single Species Method*

First, compare each stimulus induced expression profile in the *S. cerevisiae* compendium to the control expression profile. Then, search the *S. cerevisiae* genome for (i) individual genes and (ii) gene categories whose expressions differ significantly; an optimal threshold expression difference, e.g. 20%, will need to be determined. For (i), search only the intergenic space around the selected genes for correlations between motifs. Apply the same procedure for (ii), but search around all genes in the selected gene category.

(2) *Multiple Species Method*

A preliminary for this CGA approach is to compile gene expression profile compendiums for *S. paradoxus*, *S. mikatae*, and *S. bayanus* according to the same stimuli used by Hughes et al. in the *S. cerevisiae* compendium.[3]

First, compare the *S. cerevisiae* expression profile induced by stimulus X with the expression profile induced by stimulus X of each other species. There will then be two options. (i) The profiles show differences: the differential expression may be accounted for my differential motif patterns. So, apply the single species method of searching for motif correlations in the *S. cerevisiae* genome only in the intergenic space around genes and gene categories that show significant differential expression. (ii) The profiles show no difference: this implies that significant differences in motif patterns between species are irrelevant to the pathways invoked in response to the given stimulus. Also, comparing the *S. cerevisiae* profile induced by X with the control *S. cerevisiae* profile will give the differential gene expression due to these invoked pathways. So, in *S. cerevisiae*, search the intergenic space around the genes and gene categories that show both differential expression (between $\text{Profile}_{\text{Control}}$ and $\text{Profile}_{\text{X}}$) and significant differences in motif patterns (between species). Since the differences in these particular motif patterns do not alter gene expression between species, it should follow that the pattern differences do not imply additional motif interaction. It is difficult to derive a priori conclusions in this case, since the possibilities are many. But this gained negative information would definitely be useful in elucidating the motif interaction network.

(3) *Determining Motif Interaction Dependence on Frequency and Proximity in Intergenic Space*

Ideally, the ultimate goal of this study would be to produce a motif interaction network that gives not only motif interaction partners, but also gives quantitative information on the degree to which the motif interactions depend on motif frequency and proximity in intergenic space. But this latter aspect is terribly complex. In fact, bioinformatics techniques might only be fruitful after more elementary knowledge is acquired. Thus, the best initial approach might be to perform *in vitro* experiments using artificial DNA sequences with interacting motifs varied by frequency and proximity. Relevant transcription factors and the standard transcriptional machinery would also be added to reconstitute the transcriptional system. Transcription rates of the varying DNA sequences would then be measured and compared, hopefully to yield useful dependence information on the frequency and proximity parameters.

<u>Citations:</u>

1. Brent MR, Guigo R. "Recent Advances in Gene Structure Prediction." Curr Opin Struct Biol. 2004 Jun; 14(3):264-72.
2. Kellis M, Paterson N. "Sequencing and comparison of Yeast Species to Identify Genes and Regulatory Elements." Nature. 2003 May 15; Vol 423:241-54.
3. Hughes T, Marton MJ. "Functional Discovery Via a Compendium of Expression Profiles." Cell. 2000 Jul 7; Vol 102:109-26.

Peter Starr
Bioinformatics Final Project
12/10/05