Laura Mustavich

Mark Gerstein

Genomics and Bioinformatics

10 December 2005

Optimal Species Choice for Comparative Genomics: A Review and Extension of

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

Summary

Background

The goal of this paper was primarily to identify the genes, regulatory elements, and the function of these regulatory elements, of *Saccharomyces cerevisiae*, by comparing its genome to those of related organisms. Their methods are based on the premise that variation at the nucleotide level is much lower in genic, rather than intergenic regions. Therefore, genes can be identified by those regions that possess a much higher degree of conservation across genomes, than expected by chance. The accuracy is contingent on how wisely the species are chosen. They must be related closely enough, so that synteny is conserved, but distant enough to facilitate recognition of functional elements and genes. If the species are too closely related, non-functional sites will not have had enough time to undergo enough genetic drift to distinguish themselves from functional sites, resulting in many false positives. If the species are too distant, however, there will not be enough orthology to establish these regions as functional or not, resulting in many false negatives. In this case, they compare the genome of *S. cerevisiae,* as found in the *Saccharamyces* Genome Database, to those of *S.*

mikatae, S. paradoxus, and *S. bayanus,* all of which diverged from *S. cerevisiae* 5-20 million years ago (Kellis 241).

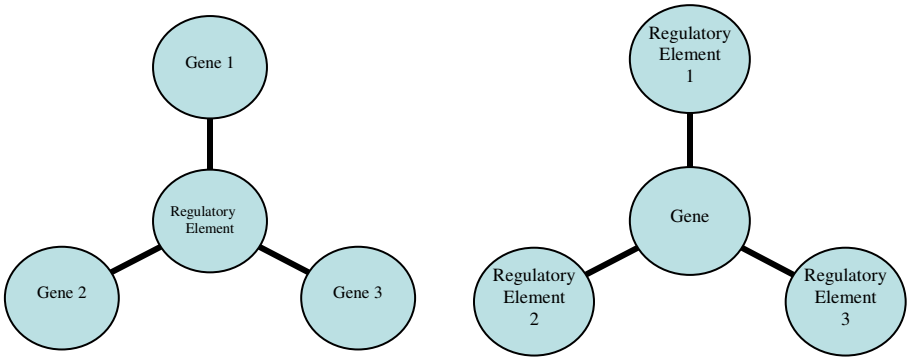
Methods

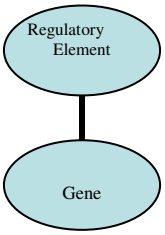
They identified genes of *S. cerevisiae* using the Reading Frame Conservation (RFC) test, developed by them to classify *S. cerevisiae* ORFs as biologically meaningful or meaningless, by the proportion of conservation among the species. They first identified the ORFs of *S. cerevisiae,* defined as sequences greater than 50 amino acids long, beginning with a start codon, and ending with a stop codon. They then aligned the genome of *S. cerevisiae* separately, to each of the other three species; first globally, to find orthologous ORFs, then locally, to align the ORFs. Since the RFC scores followed a bimodal distribution, a threshold was defined for each species. Subsequently, each species "voted" on whether an ORF was a valid gene, or abstained if it did not contain an orthologous region (Kellis 243).

They employed two different methods to identify regulatory elements: genome-wide and category-based. For the first method, they listed all possible mini-motifs of the form $XYZn_{(0-21)}UVW$. They determined which ones were conserved by checking if each mini-motif followed at least one of the following conservation criteria: 1) the mini-motif shows significantly high conservation rate in intergenic regions, 2) there is significantly higher conservation of the mini-motif in intergenic, rather than genic regions, and 3) there are significantly different conservation rates of the mini-motif between upstream and downstream regions. Conserved mini-motifs are connected to form extended motifs, which are clustered if they tend to occur in the same intergenic regions. These motifs are assessed by the Motif Conservation Score (MCS), which evaluates the genome-wide

conservation rate of the motif in intergenic regions, measured in standard deviations above the rate for comparable control motifs. Those with MCS greater or equal to 4 are retained (Kellis 248-249).

The category-based method was similar, but aimed to discover regulatory elements by searching for conservation of the motif within individual gene categories, of which there are 318 yeast gene categories, rather than genome-wide. This was accomplished by adding one more constraint to the original conservation criteria; conserved mini-motifs are those enriched in the intergenic regions of genes in the category. Finally, the functions of these regulatory elements were inferred by looking at adjacent genes. They were also able to refine their definitions of gene structure, identify rapidly and slowly evolving genes, and determine which regulatory elements act by combinatorial control in the process (Kellis 252).





Combinatorial Control
In many cases, one regulatory element regulates one gene. However, there is not always a one-to-one relationship. In some cases, many regulatory elements work together to control one gene, while in others, a single regulatory element controls multiple genes.

Commentary

The main concept underlying their methods is logical. Their methods also seem very powerful. In their test for identification of genes for instance, 96% of the ORFs were rejected, indicating high sensitivity, while 3% were not rejected and likely true ORFs. The remaining 1% were not rejected and probable false positives, the low number indicating high specificity (Kellis 243).

Their method also had many weaknesses. Though they triumphed over their many rejections, claiming this indicated high sensitivity, these negatives might not all be true. They justified them by mentioning there is no experimental evidence for the rejected ORFs. However, just because there is no experimental evidence for an ORF, does not mean it is not real. It could be that an ORF has simply never been studied. Therefore, their sensitivity is probably over-estimated. They even admit that their method does not have sufficient power to discriminate between genic and intergenic regions for ORFs encoding proteins less than 50 amino acids long, and does not detect rapidly evolving genes, such as those located in telomeres (Kellis 245). They did mention, however, that the rapidly evolving gene YBR184W passed the RFC test after performing multiple alignment instead of separate single alignments (Kellis 247). Further research should be done to determine whether multiple alignment is more powerful than single alignments overall in these methods.

They were very vague about how to select the species used in the test, however, which is a grave drawback since the choice of species is crucial to the validity of the analysis. They mentioned that the branch length between the species they used ranged from 0.23 to 0.55, for a total branch length of 0.83, a signal-to-noise ratio of 2, and the

probability of nucleotide identity across all 4 species in non-coding regions of 0.49 (Kellis 253). There was no discussion, however, of whether this choice is optimal, and how they arrived at these particular species. Additionally, I am doubtful of whether three other species is enough for such analysis. Despite varying degrees of evolutionary distance between the species used, they are treated equally in the analysis. It seems that conservation between the primary species and a distant species is more significant than conservation with more closely related species, and should consequently be more emphasized in the test. Therefore, further research should be done to determine the optimal number of species needed for the analysis, as well as the optimal range of evolutionary distance among the species. A system could also be developed to weight the votes of each species according to their evolutionary distance to the primary species.

Extension

Problem

Despite the importance of species selection in comparative analysis, little research has been done on how to optimally choose species. While there is a general consensus that the ability to extract useful information from genomes in comparative genomics studies depends on a balance between too little and too much divergence from the target species, the guidelines that govern this decision still remain subjective. It is apparent that Kellis et al. mainly based their decision to use *S. mikatae, S. paradoxus,* and *S. bayanus* in their study on the guesswork of Cliften et al, who thought *S. cerevisiae* should be compared with at least one species from the different subgroups of the genus *Saccharomyces*: the senso stricto species, which are physiologically similar to *S. cerevisiae*, and the senso lato and petite-negative species, which are quite different from

S. cerevisiae, thereby achieving the balance between little and much divergence (Cliften 1175). This decision process, however, is not based on any concrete rules.

What has been done so far

Though no method has been developed to choose the optimal species for such analysis, a few abstract models have at least been proposed. McAuliffe et al. have designed a hypothesis test that uses a statistic they call the fully observed symmetric star topology (FOSST) likelihood-ratio statistic, based on the error rates for detecting and overlooking conservation at a single orthologous site (McAullife 7900). Cooper et al. have proposed a mathematical model where analytical strength increases as the total neutral branch length of the phylogeny $\left(\sum_i d_i\right)$, since the probability that a neutral site would be misclassified as conserved is approximately $e^{-\sum_i d_i}$ (Eddy 0095). Their model assumes that conserved sites are invariant however, which greatly reduces their sensitivity. Like McAuliffe et al., Sean R. Eddy has proposed a model which assumes a Jukes-Cantor process, in which all types of base substitutions occur at the same rate (Eddy 0096). It is much simpler and intuitive, however, while much more general than the Cooper model. Its focus is to choose the optimal number of additional species, N, with the optimal evolutionary distance, D, from the target species, the branch length measured in the number of neutral substitutions per site. This statistical power is expressed by the false positive rate, FP, and the false negative rate, FN, and is given by the following equations (Eddy 0096):

$$FP = P(\leq C changes \mid neutral) = \sum_{c=0}^{C}\left[\binom{NL}{c}\left(\frac{3}{4}-\frac{3}{4}e^{-\frac{4D}{3}}\right)^{c}\left(\frac{1}{4}+\frac{3}{4}e^{-\frac{4D}{3}}\right)^{NL-c}\right]$$

$$FN = P(> Cchanges \mid conserved) = 1 - \sum_{c=0}^{C} \left[\binom{NL}{c} \left(\frac{3}{4} - \frac{3}{4} e^{-\frac{4\omega D}{3}} \right)^{c} \left(\frac{1}{4} + \frac{3}{4} e^{-\frac{4\omega D}{3}} \right)^{NL-c} \right]$$

where
L = the number of nucleotide sites in a sequence in the target genome
c = the number of changes observed relative to the target feature genome
C = threshold over which the feature is evolving at the neutral rate, conserved otherwise
ω = rate at which conserved features evolve relative to neutral features

for short evolutionary distances, we expect about
c = DNL changes in neutral features
c = ωDNL changes in conserved features

Though not surprising, the model confirms two general rules about the choice of appropriate species; all other things being constant, the required number of comparative genomes is inversely proportional to detectable feature size, and at small evolutionary distances, required genome number is inversely proportional to the neutral distance between the target genome, and each comparative species (Eddy 0101).

My project

Though this is an excellent advance, abstract models stop at actually determining the optimal species to use. Consequently, my goal is to create a web tool that tells the user which species to use, given a target species, in comparative genomics studies for maximum power in discerning genes and regulatory elements, using the Eddy model. I would first integrate various genome databases, such as the *Saccharomyces* Genome Database with an aim to include species from every sequenced genre. I would then use metropolis sampling with simulated annealing to sample over the space of genomes included in my integrated database, in order to find the set which minimizes the false positive and false negative rates. I would do this by minimizing the energy function E(X)=FP(X) + FN(X) where X is the current set of genomes, and FP and FN are as

defined by the Eddy model. I would align the genomes of X using BLAST. The variable N would be simply calculated from the set X. The evolutionary distance between the target species and the comparative species would be calculated for each genome in the set using the PAML software package, v3.13, then averaged to obtain D (Cooper 819). Testing could be done to find the optimal values for C, ω, and temperature *t*. Alternatively, these could be left as parameters for the user to decide, as they might be case-dependent. L would be input from the user, dependent on what feature they are looking for. For instance, L = 50, L = 8, and L = 1 are examples for detecting small coding exons, transcription factor binding sites, and single nucleotides, respectively (Eddy 0096). Of course, the user would input the target genome. The algorithm would begin with a random sample of genomes. The set X would subsequently change by randomly adding or removing a single genome from the set. $\Delta E = E(X+\Delta X) - E(X)$ would be calculated. The change would be accepted with probability $P(\Delta X) = 1$ if $\Delta E \leq 0$, or $P(\Delta X) = e^{-\Delta E/kt}$ if $\Delta E > 0$.

After developing my tool, I would test it on *S. cerevisiae,* using it to find a supposedly optimal set of genomes for the identification of genes and regulatory elements. After running it for a wide range of parameters C, ω, and *t*, to obtain many sets, I would test them by using them in the comparative genomics method employed by Kellis et al. If any of my sets gave results as good, or better, than those obtained by Kellis et al, I would release my web tool for public use, setting the parameters C, ω, and *t*, corresponding to the best performing set as the default parameters.

Works Cited

Cliften, Paul F., LaDeana W. Hillier, Lucinda Fulton, Tina Graves, Tracie Miner, Warren R. Gish, Robert H. Waterston, and Mark Johnston. "Surveying *Saccharomyces* Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis." Genome Research 11 (2001): 1175-1186.

Cooper, Gregory M., Michael Brudno, NISC Comparative Sequencing Program, Eric D. Green, Serafim Batzoglou, and Arend Sidow. "Quantitative Estimates of Sequence Divergence for Comparative Analyses of Mammalian Genomes." Genome Research 13 (2003): 813-820.

Eddy, Sean R. "A Model of the Statistical Power of Comparative Genome Sequence Analysis." PLOS Biology 3 (2005): 0095-0102.

Kellis, Manolis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. "Sequencing and comparison of yeast species to identify genes and regulatory elements." Nature 423 (2003): 241-254.

McAuliffe, Jon D., Michael I. Jordan, and Lior Pachter. "Subtree power analysis and species selection for comparative genomics." PNAS 102 (2005): 7900-7905.