Genomics and Bioinformatics: Final Paper - 2005/12/10

Jiang Du

jiang.du@yale.edu

Sequencing and comparison of yeast species to identify genes and regulatory elements *Kellis et al. 2003*

Review

Comparative genome analysis of related species has been an important approach for identifying functional elements without previous knowledge of function. There are several reasons for this approach to be frequently used. First of all, identification of the complete genome sequences of different species remains to be imperfect, which makes the prediction based on a single genome sequence not that reliable. With the power of comparative genomics, errors in different genome sequences can be either identified or suppressed by utilizing proper computational methods. Second, functional elements within the genome sequence will stand out by the virtue of having a greater degree of conservation across related species, given that the selection of species for study is appropriate. Third, the comparative approach can also reveal the evolutionary differences among different genes across related species (for instance, species-specific genes, rapidly/slowly evolving genes), which may in turn lead to a better understanding of the function of these genes.

The paper of *Kellis et al. 2003* introduced comparative genomics into the study of yeast species to identify genes and regulatory elements. This study involved four related species: *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus*. Among all these species, *S. cerevisiae* was most thoroughly studied, with a most complete genome sequence, so the authors of the

paper decided to use its genome sequence as a basis for comparison.

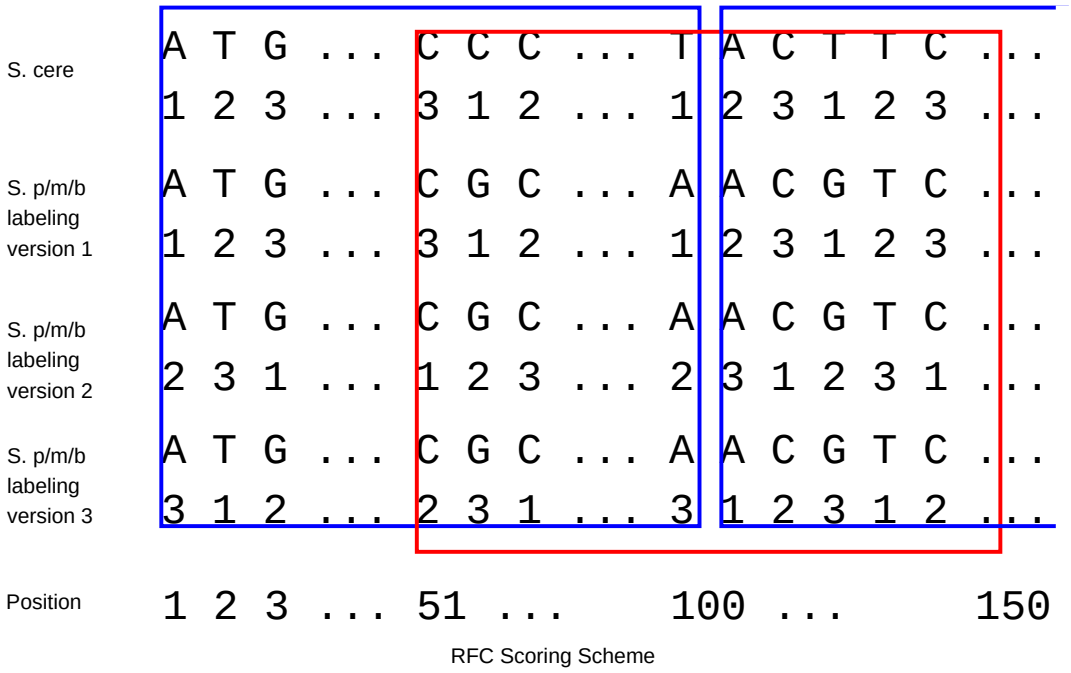
A set of studies was done, including the study of genome evolution, identification of genes, and identification of regulatory elements.

Study of genome evolution

The study of genome evolution was based on genome alignment of the four species and then the investigation of one-to-one ORF matches (which was in turn based on BLAST and graph manipulation to obtain unambiguous matches). After obtaining the ORF matching information, the genome evolution was studied both at a large scale and at the nucleotide level. Rapid structural evolution in the telomeric regions were observed at large scale comparison. As to the nucleotide level comparison, tremendous conservation of synteny was found and the overall rate of sequence divergence across the species with a remarkable contrast between intergenic and genic regions showed the great potential for further study of gene identification.

Identification of genes

In this paper, the first step of identification of genes was to apply a reading frame conservation (RFC) test to classify each ORF in *S. cerevisiae* as biologically meaningful or meaningless. The RFC score, as described in the supplementary information of the paper, was calculated by averaging the maximum conservation percentages obtained at overlapping windows of 100 nucleotides starting every 50 nucleotides, and then tallying the result of using all three species. The basic scheme of this procedure is shown in the following figure: For each window of length 100 nucleotides, a maximum label conservation score among the three different labeling versions was chosen, and the all such scores were averaged to generate a final score for this ORF with each *S. par/mik/bay*.



RFC Scoring Scheme

The authors then discovered that the RFC scores obtains showed a clearly bimodal distribution, which allowed a simple thresholding to make a tallied decision on whether to accept or reject a proposed ORF. Manual inspection was performed on the results of this RFC test and revealed satisfying sensitivity and specificity.

Rapid and slow evolution of genes was also studied in this paper, and different characteristics of species-specific genes, rapidly/slowly evolving genes were discussed.

Identification of regulatory elements

Three conservation criteria (CC1-3) were proposed based on the authors' observation about the properties of the Gal4 motif: intergenic conservation (CC1), intergenic-genic conservation (CC2), and upstream-downstream conservation (CC3). The genome-wide motif discovery was done by computing the conservation criteria score of all 45760 possible sequences of the form $XYZn_{(0-21)UVW}$, and then extending those conserved mini-motifs to construct full motifs, whose MCSs (Motif conservation score, which was based mainly on comparing the statistics of the motif to those of its random alternatives) were then calculated to reveal those motifs with comparable significance. This procedure reported 72 well-conserved motifs,

which included most known regulatory motifs, with a comparable number of new motifs.

The authors then did a further interesting research in inferring function of these discovered motifs, based on the function of the genes adjacent to conserved occurrences of the motif with know gene categories. This work, was of course, based on the common belief that there was a strong correlation between the motif and the genes adjacent.

Further identification of regulatory elements based on category information was described and some study on combinatorial control of multiple motifs were also discussed at the end of this paper.

Comments

One of the ideas that formed the basis of this paper was the proper selection of the related species to apply comparative genomics approaches. At the discussion section of the paper, the authors proposed several general principles in this issue, which were of course all very sound. However, what was not taken into consideration in that discussion was the possible impact on the whole process from imperfect genome sequencing. For instance, researchers might find a perfect candidate for comparative study (maybe based on their knowledge in other related fields in biology), but might then be presented a genome sequence with a relatively high error rate, which just could not be ignored in this case. One solution to this case was to wait until there was a better sequence or to select another candidate, but could there be any other solutions that still involved this particular candidate? One could also ask a related question, which would be: how would the errors in genome sequences affect the results of comparative genomics? This question seemed to be related to the robustness of the comparative approach when the input data were permuted.

Besides the concern of the general issue above, the details of some computational methods used in this paper might also be worth of further investigation. For example, the RFC test in this paper used a labeling and windowing approach to compute the sequence conservation, which generated satisfying results. However, one might want to ask in the first place that

what the actual meaning of sequence conservation was, and this discussion might lead to alternative computing methods with a stronger biological basis.

Anyway, this paper in general gave a thorough investigation of what could be done in comparative genomics and reveal new results in identifying genes and motifs in yeast species, which meant that the paper was sound and important in both bioinformatics and genomics.

Possible further work

As discussed above, one possible further work related to this paper is to investigate the performance of the overall comparative approach with perturbed input data. The perturbation of input data would involve two different kinds: the inclusion of inappropriate species for comparison and the use of a genome sequence data with a high error rate. The importance of this study is somewhat obvious: if the comparative approach is shown to be robust even when exposed to a certain high level of perturbation, researchers can then have more confidence when dealing with perturbed data in the future. However, if this comparative approach is vulnerable even to some very small perturbation, either a strict requirement of input data should be satisfied in future research of comparative genomics, or alternative methods should be considered. Another extreme case would be that similar result would be obtained even with dramatically perturbed data, which will lead to a second thought of the validness of the comparative approach. This is however extremely unlikely to happen, so it will not be the focus of the following discussion. Anyway, it seems that the most possible result would not fall into these extreme cases, and would reveal that the comparative approach would be robust within a small range of perturbation. In this case some important thresholds could be learned from this study, and could be used as guidelines for future research.

The subject of study

The main goal of this further work is to study to performance of comparative approach on perturbed data. A question that follows would be how to establish this study: should an

abstract model for comparative operations be defined and investigated, or should the study focus on certain species? Since the paper discussed in the previous sections was about yeast species, which is an well-studied subject compared to many other species, it is reasonable to do the further work base on yeast species as well. As to abstract models for comparative operations, it makes sense to pick one operation to do a detailed study first rather than considering all the possible operations at once.

Perturbed data

As discussed previously, data perturbation in this scenario involves perturbation of input species and perturbation of input genome sequences. However, one might raise two questions on this statement: First, the perturbation of input species would anyway result in a different genome sequence, so why not just consider a single perturbation case of genome sequence change? Second, how to get the perturbed data in the first place? In other words, how to generate/simulate a biologically reasonable/meaningful perturbed dataset?

The answer to the first question lies in the different characteristics of the two kinds of perturbation. A change of species will/may change the whole sequence dramatically, while a perturbation of an existing genome sequence tends to be more subtle. The concept of two kinds of perturbation also reflects the idea of biologically meaningful perturbation: while we can simulate a dramatically changed genome sequence solely based on manipulating an existing sequence, this may not as meaningful as taking the sequence from other species in the sense of biology (this statement may, however, lead to a debate on the meaning of biologically meaningful, which is another interesting topic and beyond the scope of this discussion).

Having the first question answered (to some extent, at least), we will now focus on how to generate reasonable perturbed input data for both types of perturbation.

Perturbation in species selection Usually the focus of the species to be studies would be a well-studied organism with a relatively well established genome sequence, such as *S. cerevisiae*. What might go "wrong" would be the choice of other related species. Species that are too close to or too far away from the original one would be "wrong" selections, and would lead to inappropriate results in comparative studies. However, at this moment we don't know what are exactly the "wrong" choices since we don't have the results yet.

One approach we can try is to select an organism that is relatively close to *S. cerevisiae* but does not fall into the same narrow taxon as *S. par/mik/bay*. By doing so we are simulating the case of a small perturbation in species selection, and this can actually be done by looking at the existing phylogenetic tree/network to select a reasonable candidate. The genome length of this candidate will possibly be similar as *S. cerevisiae* and there is a good chance that it will allow a meaningful alignment in the first step of the comparative approach.

However, as we go further away in the phylogenetic tree to select candidates, the sequence obtained might be very different from *S. cerevisiae* and we may not even be able to get a meaning alignment. There are several solutions/workarounds. On one hand, we can stop at a point that we set beforehead, where further selection would make the whole comparative approach obviously meaningless and will (hopefully in most cases) be noticed by human researchers. On the other hand, when it is the case that the selected genome sequence is too long (for example, the human genome) compared to *S. cerevisiae*, we may modify the sequence (most probably by deletion) so that it is comparable to *S. cerevisiae*. This is applicable since we can still find orthologs of the yeast ORF in human genome, and the comparison of surrounding regions may still be scientifically meaningful.

Perturbation in genome sequence Actually the paper in previous discussion already reveal the robustness of the comparative approach to some extent, since the sequence used there was not perfect in the first place. However, a relatively stronger perturbation on the genome sequence could not be achieved without simulation.

There are already quite a few literature out there discussing how this sort of simulation could be done in a (supposedly) meaningful manner. For example, according to the error rate to achieve, a number of locations in the genome sequence can be chosen in a random fashion, and certain perturbations (insertion, deletion, translocation, and simple swappings) can then be applied to the regions in these locations to generate a a genome sequence with the expected error rate. Despite the soundness of this approach, one might ask whether this represents the actual errors that may occur during sequencing. An alternative approach could be to study the characteristics of the actual errors in sequencing first (this may be done based on an investigation of different sequencing versions of a single genome) and then try to simulate the errors in a similar way.

Comparative approaches

A remaining decision to make is to choose a certain comparative operation to study in the first stage. The principle in choosing such an operation would be to chose a simple but still meaningful one, as in most of other studies. A good candidate in this case can be the RFC test, whose performance can be relatively easily assessed by applying it on named ORFs and control sequences of intergenic sequences. The sensitivity and specificity can be recorded for different input data and further analysis can be done based on the records.

Another reason for choosing the RFC test as a study candidate is that pure statistical analysis can also be done on its scoring scheme, and theoretical results of its performance on random data can be calculated and compared to the results based on actual data for us to better understand the whole problem.

What to expect

As discussed previously, although this study is to investigate the performance of comparative approaches on perturbed data, no breath-taking result is expected (although it is welcome...). Rather, a better understanding of the circumstances where the comparative approach is

suitable and its performance is the ultimate goal of this study.

References

1. Kellis, M. et al. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-254 (2003).

2. Goffeau, A. et al. Life with 6000 genes. *Science* **274**, 546, 563-567 (1996).

3. Altschul, S. F. et al. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).

4. Batzoglou, S. et al. ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12**, 177-89 (2002).

5. Jaffe, D. B. et al. Whole-genome sequence assembly for Mammalian genomes: arachne 2. *Genome Res.* **13**, 91-6 (2003).

6. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16C23 (2000).

7. Pennacchio, L. A. et al. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2**, 100-109 (2001).

8. Meyer, I. M. et al. Gene structure conservation aids similarity based gene prediction, *Nucleic Acids Res.* **32**, 2, 776-783 (2004).

9. Blanchette, M. et al. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12**, 739-748 (2002).