Poster A-27

Prediction of Non-Coding RNAs Using Drosophila Whole-Genome Alignments

Authors:
Yuri Bendana (*Bioengineering Graduate Group, UC Berkeley/UCSF*)
Ian Holmes (*Bioengineering, UC Berkeley*)

Short Abstract: We investigate the feasibility of using a whole-genome alignment of Drosophila species for predicting functionally conserved features such as non-coding RNAs de novo and by annotation transfer.

Long Abstract:

The recent sequencing of multiple species of Drosophila provides the opportunity to perform investigations into the feasibility of using whole-genome alignments as a tool for predicting gene elements. It is commonly believed that evolution conserves the most essential parts of the genomes across species. By using comparative genomics, these functionally important regions can be identified from the more variable regions.

In particular, non-coding RNA perform important regulatory functions by controlling gene expression. For example, antisense RNAs bind their target mRNAs to inhibit their translation. However, it may be difficult to identify non-coding RNAs by sequence alone, since although their structure is important for their function, their sequence may vary. Non-coding RNAs should display covariation in their mutations because of basepairing interactions. For this reason, it is believed that multiple alignments of closely related species will display the evolutionary signals of non-coding RNAs more clearly.

Some methods allow for the simultaneous pairwise alignment of sequences and their annotation. However, for multiple genomes these methods usually prove to be too computationally expensive. An alternative is to first align the genomes using a tool such as MAVID, which performs a constrained multiple alignment based on known protein-coding regions. Then this alignment is used to either apply methods for de novo prediction of features or to transfer known annotations from a reference species to the other species.

For de novo prediction, we use the Xfold and PFOLD tools to predict the consensus secondary structure of an RNA alignment. We use the likelihood of the secondary structure as a measure of how probable the alignment contains a non-coding RNA. These tools use Stochastic Context-Free Grammars (SCFG) to model the long-range basepairing interactions of a folded RNA. They also model evolutionary correlations via a phylogenetic tree that relates the sequences in the alignment. The specific innovation of Xfold is the ability to 'train' (i.e. estimate maximum-likelihood rate and probability parameters) for user-specified phylo-SCFGs, allowing rapid evaluation and prototyping of new models. We compare the output of Xfold and PFOLD in a complete scan of a MAVID alignment of the Drosophila genomes. In addition we compare their output to RNAz, a tool that in addition to predicting a consensus RNA structure also provides a measure of thermodynamic stability.

We also describe our work in transferring annotations from a reference species to the other species in the alignment. In this case, we start with the annotated non-coding RNAs for D.melanogaster from Flybase. We extract the regions in the MAVID alignment containing the known non-coding RNAs. In cases where there is a high percent identity with the reference sequence, the annotation can be transferred immedidately. Otherwise, Xfold and other tools are used to predict the likelihood of RNA secondary structure and whether the annotation can be transferred.

Finally, similar approaches to applying SCFGs to whole-genome alignments are currently being explored in areas such as studying covariation in TIR transposon families. These will be described briefly.