DrosOCB: a high resolution map of conserved non coding sequences in Drosophila

ID - 194

Martignetti Loredana[1], Michele Caselle[1], Bernard Jacq[2], Carl Herrmann[2]
[1]Department of Theoretical Physics, University of Torino
[2]Institut de Biologie du Developpement de Marseille-Luminy

Motivation

The functional annotation of eukaryotic DNA sequences represents a great challenge in post-genomic biological research. A great aid to functional annotation of genome sequences is provided by comparative genomics methods which, since a few years, have been extended also to non coding DNA regions. However, comparison of non coding sequences requires new algorithms and strategies to take into account the different evolutionary mechanisms affecting regulatory sequences. Here, we present an novel large scale alignment strategy which aims at drawing a precise map of conserved non-coding regions between genomes, even when these regions have undergone small scale rearrangement events. Our procedure is optimized to take into account the great plasticity of non coding DNA, such as shuffling and sequence variability of binding sites within functional modules, low scale translocations, inversions and duplications.

Methods

The recent availability of 12 Drosophila species sequencing and annotation offers a complete and reliable genomic dataset for developing and testing methods for comparative genomics of non coding DNA.We used a 'gene-centric' approch, in that it starts with a list of orthologous genes between two species, and applies a local alignment algorithm to the corresponding flanking intergenic regions and intronic regions of these orthologous pairs.

For each Drosophila species took in exam, we compile a list of orthologous genes with D. Melanogaster, according to the '12 genomes project' data.

Considering each locus in our list, according to the mentioned genome annotation, we selected the whole repeated-masked sequences containing the transcriptional unit and the corresponding flanking intergenic regions up to the preceding and the following gene.In this way, our selection is not constrained from the syntenic order of the orthologous genes, which is relevant when we compare species very distant in the evolutionary tree.

Local pairwise alignments between related sequences was performed using CHAOS, which is an heuristic alignment algorithm with some peculiar features optimized for large non coding DNA sequences. CHAOS works chaining small words which match between the two input sequences. Differently from BLAST, it is a double seed technique and it allows some degeneracy in seeds. It chains toghether seeds that are closer that a maximum distance and it returns the highest scoring chains, according to a standard Needleman-Wunsch metric. Hence, it is able to identify conserved blocks rearranged in non-colinear order or in a reverse order with a very high resolution. On the other hand, it is able to rapidly align large sequences with a better specificity than purely local aligners, thanks to the double seed technique.We choose two different sets of parameters in CHAOS and build a conservative and a more sensitive version of our alignments.

We applied the described procedure to each list of ortholgous genes between D. Melanogaster and seven other Drosophila species, providing a provisional data browser at: http://139.124.62.227/~carl/UCSC/TrackMaker.php.

Results

We obtained a genome-wide high resolution map of D. Melanogaster compared to other seven drosophila species.

According to this map, we can extimate conservation features of Drosophila genome at large scale: the percentage of conservation of intronic and intergenic genome, the rate of low scale rearrangement events, as inversions and reshuffling. Interestingly, we observe numerous small scale rearrangement events, such as local inversions, duplications and translocations, which are not observable in the whole genome alignments currently available. For example, about 15% of the conserved blocks have been obtained aligning the orthologous regions on opposite strands, indicating small scale inversions. Moreover, because we allow 1-n relationships between blocks in both aligned species, we immediately spot duplication events, like the duplication of the tRNA gene K5:84ABd of D.melanogaster in D. pseudoobscura.

This catalog of non-coding conserved blocks will constitute the starting point for several investigations, related to the evolution of conserved non-coding regions in the drosophila or the discovery of cis-regulatory regions.

Email: martigne@to.infn.it