RNA meets promoter
(The story of Old, New, Borrowed and Blue)

Wolfgang Otto

Chair in Bioinformatics, University of Leipzig

Herbstseminar, October 2009
Motivation

- background: ncRNA annotation within the worms (in this presentation only the Caenorhabditis, P. pacificus has slightly different promoters while B. malayi, M. haplanaria and M. incognita acrita have no recognisable promoter elements)
- goal: find new members of the ncRNA families (and maybe genes of unknown families)
Idea

- find ncRNA genes by their promoters
  1. analyse promoters of known worm ncRNA genes
  2. search for similar motifs in worm genomes
  3. checking DNA downstream of promoter like regions
Promoter Analyses

- ncRNAs are transcribed by polymerase II and III
  - extracted 100nt upstream of U3 snoRNA, 7SK RNA, SL1 and SL2 RNA, SmY RNA, U1, U2, U4 and U5 snRNA (pol II) and 100nt upstream of RNase MRP, RNase P, sbRNA and U6 snRNA (pol III)
  - align extracted sequences based on PSE-B box (included in every promoter sequence)
  - sequences without promoter sequences were used to identify pseudo-genes
Global Promoter Search

- used alignment to create \texttt{fragrep} pattern for each kind of promoter sequence

- remarkable: two kinds of PSE-B boxes in pol II promoter
  - pol II (a) U3 snoRNA, 7SK RNA, SL1 RNA
  - pol II (b) SL2 RNA, SmY RNA, U1, U2, U4 and U5 snRNA
Global Promoter Search

- pattern search with `fragrep` and relative variable scores $\Rightarrow$ high number of candidates
- reduce number of false positive hits
  - use mismatch-based matrix similarity score ($mmS$)
  - shuffle genomes and repeat `fragrep` search, hits are 100% false positive
  - define cutoff $mmS$ where only 1% of the false positive hits remain
  - use this cutoff $mmS$ for original hits
The Story of Old, New, Borrowed and Blue

- remaining hits are assumed to be real promoter sequences
  1. adjust hits of putative ncRNAs, used for the pattern generation \([\Rightarrow\text{old and borrowed}]\)
  2. look for new members of known families \([\Rightarrow\text{new}]\)
      - generate fasta db consisting of all known ncRNAs
      - blast search for downstream sequence of the hits (100nt) in the fasta db
  3. check remaining hits for ncRNA genes of unknown families \([\Rightarrow\text{blue}]\)
The Blue Part

- extract 70nt downstream of each promoter hit
- create clusters of similar sequences with blastclust
- remove all clusters that contain only sequences from one species
- check remaining inter-species clusters for potential ncRNA genes (yet only UCSC genome browser)
- ⇒11 potential new ncRNA families
Cluster 1 (c000124, Pol III)

- chrI: 8,245,471-8,245,683 (-); \( MFE = -14.50 \text{kcal/mol} \)
- EST, GB: BJ118936.1 (unpublished oligo-capped cDNA library, L1 stage)
Cluster 2 (c000147, Pol III)

- chrII:5,599,660-5,599,872(+); \( MFE = -21.00\text{ kcal/mol} \)
Cluster 3 (h0000010, Pol III)

- chrII:14,635,601-14,636,068(-)