Eithon Cadag
Lecture 15, 11/17/04
CSE527

Notes on GenScan

- Generalized HMM - transition probabilities based on sequences, not single nucleotides
  - P(sequence | model)
- Length distribution
  - introns - geometric (self-loop)
  - terminal exons - modeled empirically
- Submodels
  - properties vary by G+C content
  - GHMM allows submodels based on G+C
- Features
  - Intron structure - sequences as logo representation
    `5'---exon---AG | GTAAG(donor)---intron---(acceptor)TTTxCAG | G---exon---3'`
    - conservation in exons
    - polymerase transcribes to RNA
    - Spliceosome recognizes features, splices out introns
    - donor sites can be expressed as graphical forms of WMM (0th order)
      - shows relative frequency
      - size is inverse of entropy
- Acceptor splice sites (3' end of intron) - 1st order

  3' end polypyrimidine tract, 2nd order Markov Model (average over 5 preceding positions to revise training data)
- Donor splice sites (5' end of intron)
  - poor matches at one end can be compensated for by strong matches on the other end
  - χ^2 test
    - $\chi^2 = \Sigma_i \frac{(observed_i - expectedi)^2}{expected_i}$
    - over 16 cells of 4x4 {A, C, T , G}
    - if 4x4 independent, observation-expected = 0
    - χ^2 not robust with small counts

        * long-range dependencies in donor splice sites - WMM not adequate; higher-order MM requires more training data
        * U1 RNA - interacts with proteins and splice sites
            · its sequence is roughly complementary to the donor splice site consensus sequence
            · hybridizes with transcript to find its position
    - maximal dependance decomposition model (decision tree) - rebuilds χ^2 for new subsets on remaining positions
        * generate 5' splice site
        * apply a new WMM to each node in tree - compare WMM to background model

- Summary of Burge and Karlin
    - coding DNA non-random
    - use of disparate and different models integrated together

- BK training sets
    - over-representation in single exon, highly-expressed, and moderate-sized genes
    - will only do as well as training set (won't find novel genes)

- Problems with all methods
    - pseudo-genes - non-functional look-alikes (may have a poly-A tail)
    - short ORFs
    - sequencing errors (particularly frameshifts)
    - non-coding RNA
    - overlapping genes
    - alternative splicing/polyadenylation
    - hard to identify novel genes
    - species-specific peculiarities

- Other ideas
    - database search - look for similar-looking regions that are known coding regions
    - comparative genomics - comparing regions between related organisms to find commonly-conserved areas

Pre-mRNA secondary structyre predictions aids splice site recognition

- pre-mRNA - direct transcript of genome (no splicing, polyadenylation)

- hypothesis: secondary structure has important information in addition to primary sequence information
- 1st order WAM/MM (calculate log likelihood ratio)
- Secondary structure statistics
  - optimal folding energy - calculate stability of folded versus unfolded
  - max helix score - 3+ paired consecutive bases nearby
    * calculate $P_{HStart,x}$ and $P_{HEnd,x}$
    * $maxhelix_i = max(P_{HStart,x}, P_{HEnd,x})$
      $x\epsilon(i-5, i+5)$
    * highest probability that a helix will form nearby
  - neighbor pairing correlation - change pre-mRNA alphabet from nucleotides to structural symbols
    * O - unpaired base
    * P - paired base
    * S - paired stacked base