RNA multiple alignment

Project

- Please email a 3 page description by Thursday. It should include the definition of the problem.
- Each proposal will be reviewed by another classmate. It must be clear enough to a non-specialist.
- It should contain
  - Problem definition, motivation
  - Proposed methodology
  - A test plan, and some test data sets

RNA multiple alignments

- Why should we compute multiple (structural) alignments for RNA?

Structural Alignment

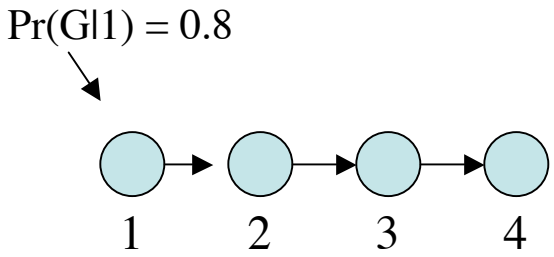
```
X07545      ..ACCCGGC.CAUA...GUGGCCG.GGCAA.CAC.CCGG.U.C..UCGUU
M21086      ..ACCCGGC.CAUA...GCGGCCG.GGCAA.CAC.CCGG.A.C..UCAUG
X05870      ..ACCCGGC.CACA...GUGAGCG.GGCAA.CAC.CCGG.A.C..UCAUU
U05019      ..ACCCGGU.CAUA...GUGAGCG.GGUAA.CAC.CCGG.A.C..UCGUU
M16530      ..ACCCGGC.AAUA...GGCGCCGGUGCUA.CGC.CCGG.U.C..UCUUC
X01588      ..ACCCGGU.CACA...GUGAGCG.GGCAA.CAC.CCGG.A.C..UCAUU
AF034619    ...GGCGGC.CACA...GCGGUGG.GGUUGCCUC.CCGU.A.C..CCAUC
L27170     AGUGGUGGC.CAUA...UCGGCGG.GGUUC.CUCCCCGU.A.C..CCAUC
X05532     AGGAACGGC.CAUA...CCACGUC.GAUCG.CAC.CACA.U.C..CCGUC
#=GC       <<<<<<<<<<.......<<.<<<<.<...<.<...<<<<.<.<.......
```

Conserved sequences, and conserved structure are more apparent in multiple alignments.

Computing Structural Alignments

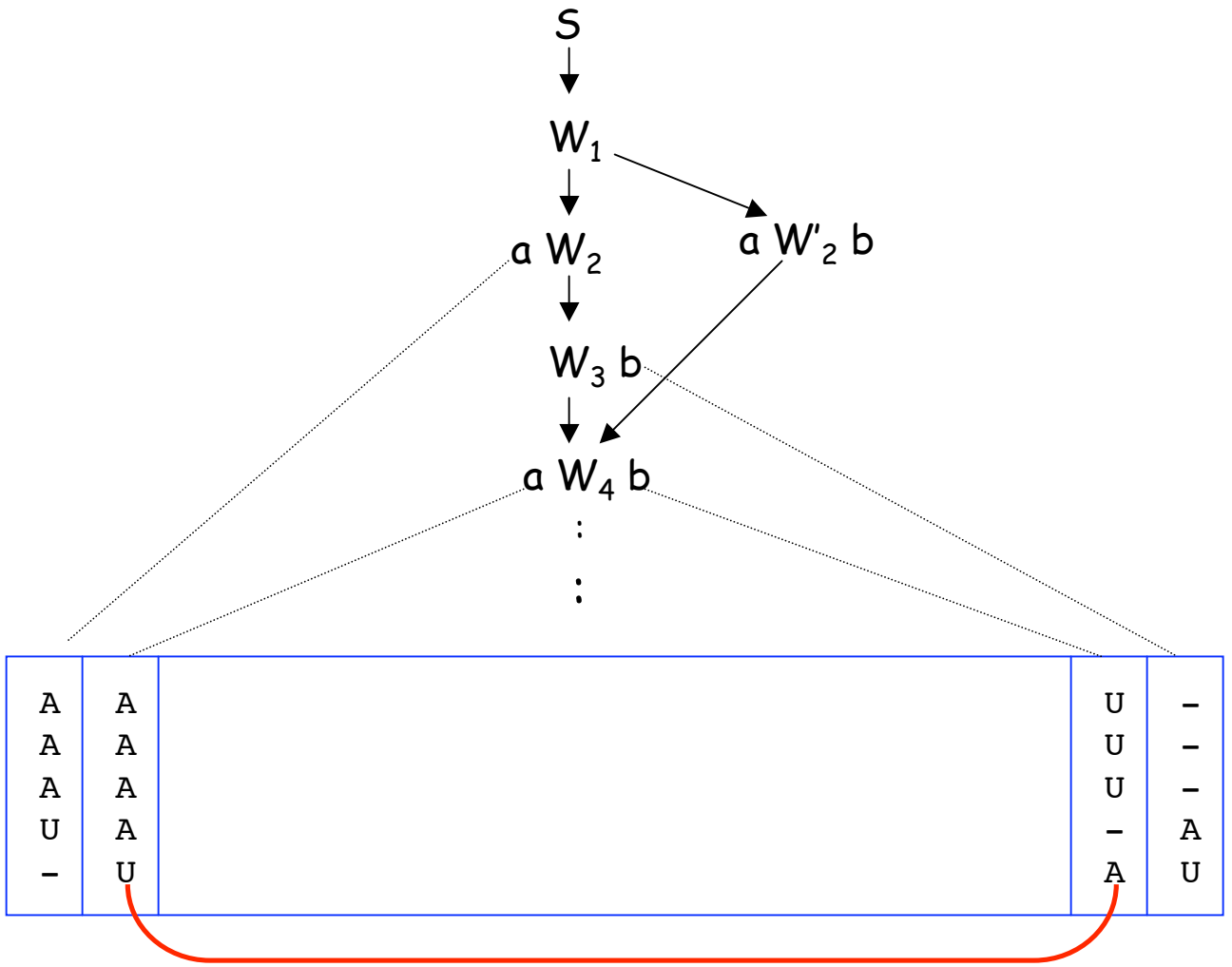
G	U	G	G	C	C	G
G	C	G	G	C	C	G
G	U	G	A	G	C	G
G	U	G	A	G	C	
G	G	C	G	C	C	G
G	U	G	A	G	C	G
G	C	G	G	U	G	G
U	C	G	G	C	G	G
C	C	A	C	G	U	C

1 2 3



- Analogy: In sequence alignment, the score for aligning a column is position independent.
- In profiles, or HMMs, position specific scoring is used to distinguish conserved positions from non-conserved positions
- Similar ideas can be used for RNA.

Covariance models=RNA profiles



Aligning a sequence to a covariance model

- We align each node of the covariance model (it is tree like, but may be a graph).
- The alignment score follows the same recurrence as in Lecture 7, but with position specific probabilities.
- Example:
  - $A[W_i,(i,j)] = -\log (Pr[W_i \rightarrow s[i]\ W_j\ s[j]\ ) + A[W_j,(i+1,j-1)]$
- If we wish to compute the probability that a sequence belongs to a family, we compute the total likelihood (sum over all probabilities)
- If we wish to compute the structure of an unknown sequence by comparison to a covariance model, we compute the max likelihood parse in this graph.

Covariance models and ncRNA discovery

- Given a family of ncRNA sequences, scan a genomic sequence with a covariance model and retrieve all high scoring sub-sequences.
- This is the most common method, but it is expensive.
- Assume covariance model has m states, and the substring has at most n symbols, and the database has L symbols.
- Alignment cost = $O(n^2m_1+n^3m_2)$
- Total time =?

Computing covariance models

- If we are given a CM, a multiple structural alignment is 'easy'.
  - In turn, align each sequence to the CM.
- If we are given a multiple alignment, computing the covariance model is easy
- For simultaneous prediction, a Bayesian iterative approach is used
  - Compute a seed alignment
  - Use the alignment to compute a CM
  - Use the CM to compute a new alignment
  - Iterate

Project

- Compute an RNA multiple alignment.
- Existing methods do not work well without good seed alignment, and require excessive hand curation.

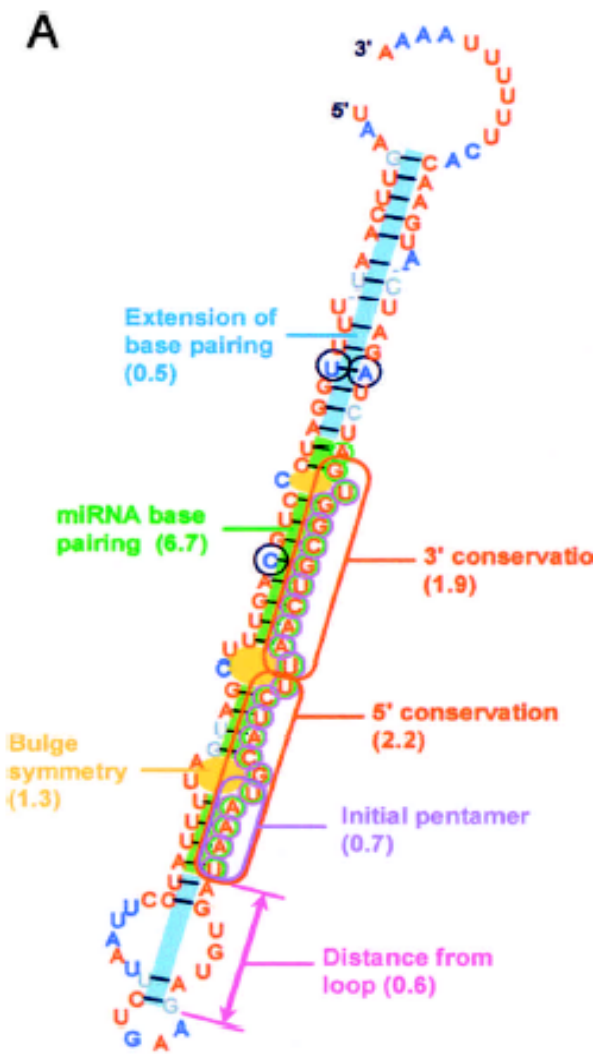
ncRNA discovery for specific families

Case study: miRNA

- dsRNA, and siRNA can be used to silence genes in mammalian tissue culture.
- miRNA is a new member of this class of endogenous interfering RNA
- RNA interference (RNAi) is a pwerful new technique to study gene function.

Case Study: miRNA

- ncRNA ~22 nt in length
- Pairs to sites within the 3' UTR, specifying translational repression.
- Similar to siRNA (involved in RNAi)
- Unlike siRNA, miRNA do not need perfect base complementarity
- No computational techniques to predict miRNA
  - Most predictions based on cloning small RNAs from size fractionated samples



miRNA (vs. siRNA)

- Derived from transcripts that form local hairpin structures.
- Sequences of the precursor, and processed miRNA is evolutionarily conserved
- Usually distinct, and distant, from other genes
- siRNA (by contrast)
  - Not evolutionarily conserved
  - Correspond to sequences of known or predicted mRNAs, transposons, or regions of heterochromatic DNA.

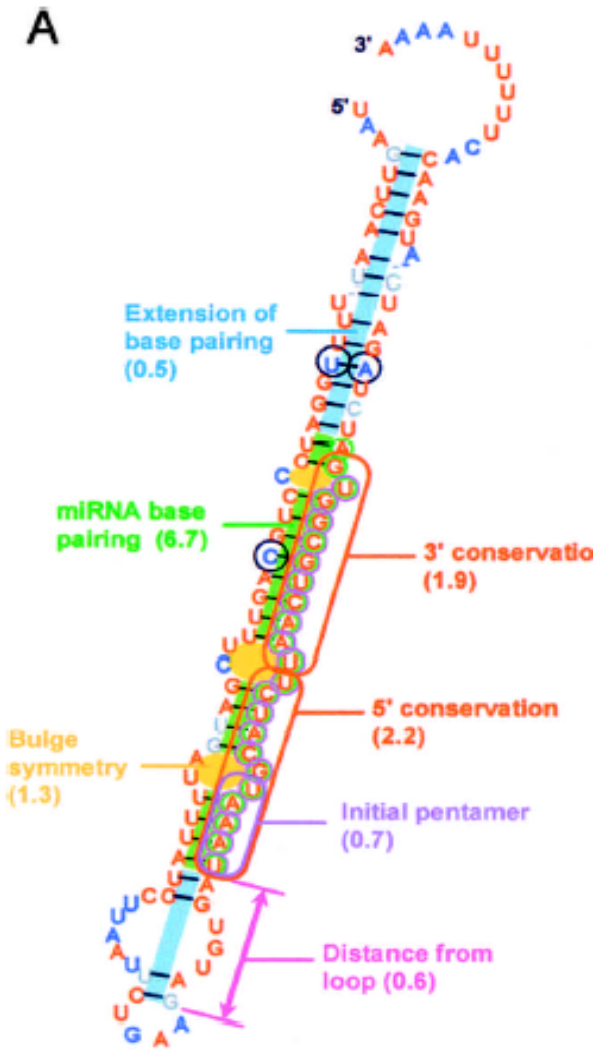
MiRscan

- Predicts miRNA
- Start with evolutionarily conserved region. Ex: *C. elegans* and *C. briggsae*
- 36000 hairpins were found (including 50/53 known miRNA).
- 50 known miRNA were used to train and score the 36000 hairpins

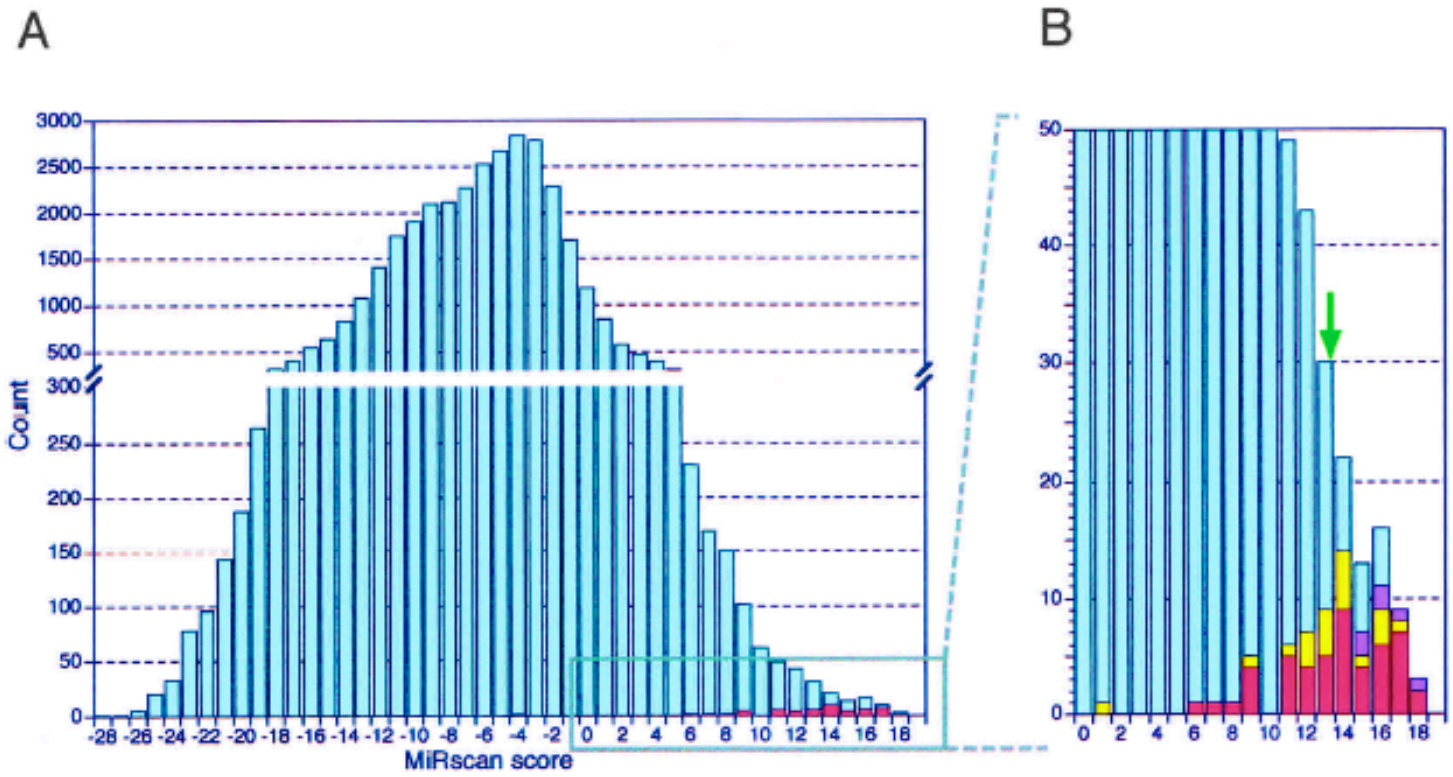
Computational identification of miRNA

- 7 features are scored
  1. miRNA base-pairing
  2. Base-pairing of the rest of the fold-back
  3. Stringent sequence conservation in the 5′ end of fold back
  4. Sequence conservation in the 3′ end of fold back
  5. Sequence bias in the first 5 bases of miRNA
  6. Tendency to form symmetric internal loops
  7. Presence of 2-9 consensus base-pairs between miRNA and terminal loop region

- Red: Conserved with *C. briggsae*
- Blue: varying residues that maintain their predicted paired or unpaired states



MiRscan scoring



- 35 previously unannotated hairpins exceeded the Median score

Molecular identification of miRNA

- Initial cloning and sequencing identified 300 clones representing 54 unique miRNA
- 10 fold scale up of the procedure identified 3423 clones as miRNA. These contain 77 distinct miRNA genes
- 77-54=23 novel miRNAs found
- 20 were scored by MiRscan (yellow). 10 were among the top 35

MiRscan results

- 35 Predictions
- 10 identified with a high throughput screen (sequencing of 3423 clones)
- 6 identified using a PCR assay.
- 4 identified as false positives PCR hybridized to larger ncRNAs
- 15 unknown
- Evolutionary conservation is important for ncRNA detection
  - >97% of all miRNA had significant conservation between *C. briggsae*, and *C. elegans*

ncRNA summary

- ncRNA, once forgotten, is increasingly important.
- Evolutionarily conserved structural elements are key to discovery and annotation.
- Special algorithms can help for specific families