An Overview of Probabilistic Methods for RNA Secondary Structure Analysis

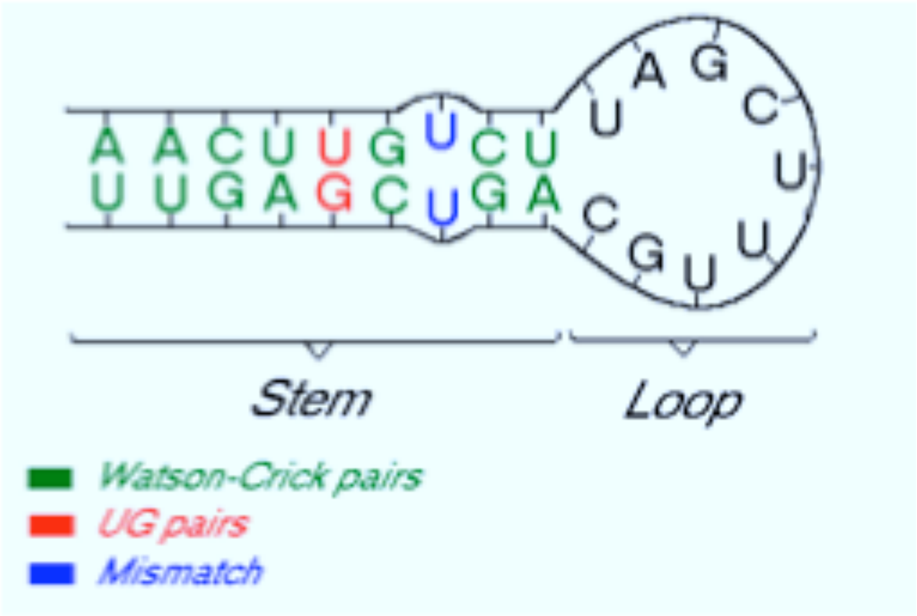
David W Richardson

CSE527 Project Presentation

12/15/2004

RNA - a quick review

- RNA’s primary structure is sequence of nucleotides (A,C,G,U)
- folds back on itself by binding stable base pairs
  - Folded structure is RNA’s *secondary structure*
- Secondary structure is the main determinant of functionality



RNA analysis

- 2 classes of RNA analysis problems:
  - Predict secondary structure of an RNA sequence
  - Create a model/profile of RNA family from a multiple alignment for:
    - Aligning new sequences to the profile
    - Searching databases for homologous RNAs that match the profile
- Solution methods based on probabilistic models of RNA secondary structure

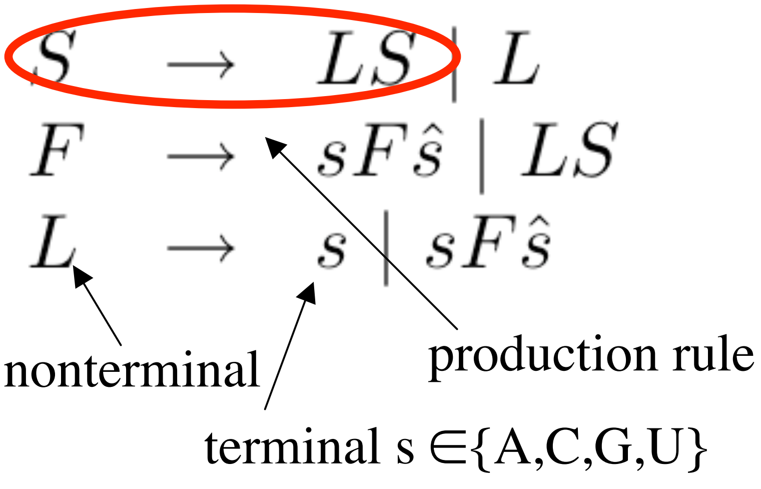
Project Outline

- Literature review of probabilistic methods:
  - Stochastic Context-Free Grammars (SCFGs)
    - SCFGs + evolutionary history (Pfold)
    - SCFGs for detecting noncoding RNAs
      - Pair-SCFGs
      - Algorithmic speedups for Pair-SCFGs
    - SCFG design considerations
  - Covariance Models
  - Brief overview of non-probabilistic methods

RNA analysis using stochastic context-free grammars

Sakakibara et al., *stochastic context-free grammars for tRNA modeling*, 1994

RNA SCFG:



$$
\begin{aligned}
S &\rightarrow LS \mid L \\
F &\rightarrow sF\hat{s} \mid LS \\
L &\rightarrow s \mid sF\hat{s}
\end{aligned}
$$

nonterminal

production rule

terminal s ∈{A,C,G,U}

$s\hat{s}$ = paired bases

Derivation/Parse-Tree of Sequence CAGUUCU from SCFG:

S --> LS --> CS --> CL --> CAFU --> CAGFCU --> … -->CAGUUCU

Key: parse trees <=> secondary structure

SCFG algorithms (DP-based)

- Secondary structure prediction
  - **CYK algorithm**
    - Given RNA sequence s and SCFG, find most likely secondary structure of s? Find most likely parse-tree of s

- Likelihood of a sequence
  - **Inside algorithm**
    - Probability that s is generated by SCFG? Similar to CYK

- Search database for homologous RNAs
  - Score subsequences using Inside or CYK
    - Log-odds or Z-scores

SCFG algorithms (DP-based)

- SCFG parameter estimation
  - **Inside-Outside algorithm**
    - EM style procedure from training sequences
    - Time cubic in length of training sequences
  - Tree-Grammer EM training algorithm
    - Faster, but needs initial structural alignments of RNAs in family

Paper’s results

- Trained 4 grammars on 1477 tRNA sequences

- Generated multiple alignments using grammars on known EMBLtRNA alignments
  - 99% base-pairs matched known alignment
  - 83% for “Part III” class of sequences (mitochondrial tRNA lacking D-domain)

- Inside algorithm generated Z-scores to discriminate tRNAs from non-tRNAs
  - Good discrimination except for Part III group

Discussion

- SCFG-based techniques are effective
- SCFGs don’t model introns, insertions and deletions
  - Necessary for real-life profiles for DNA-level database searches
- Paper doesn’t explicitly discuss database search methods

RNA analysis using covariance models (CMs)

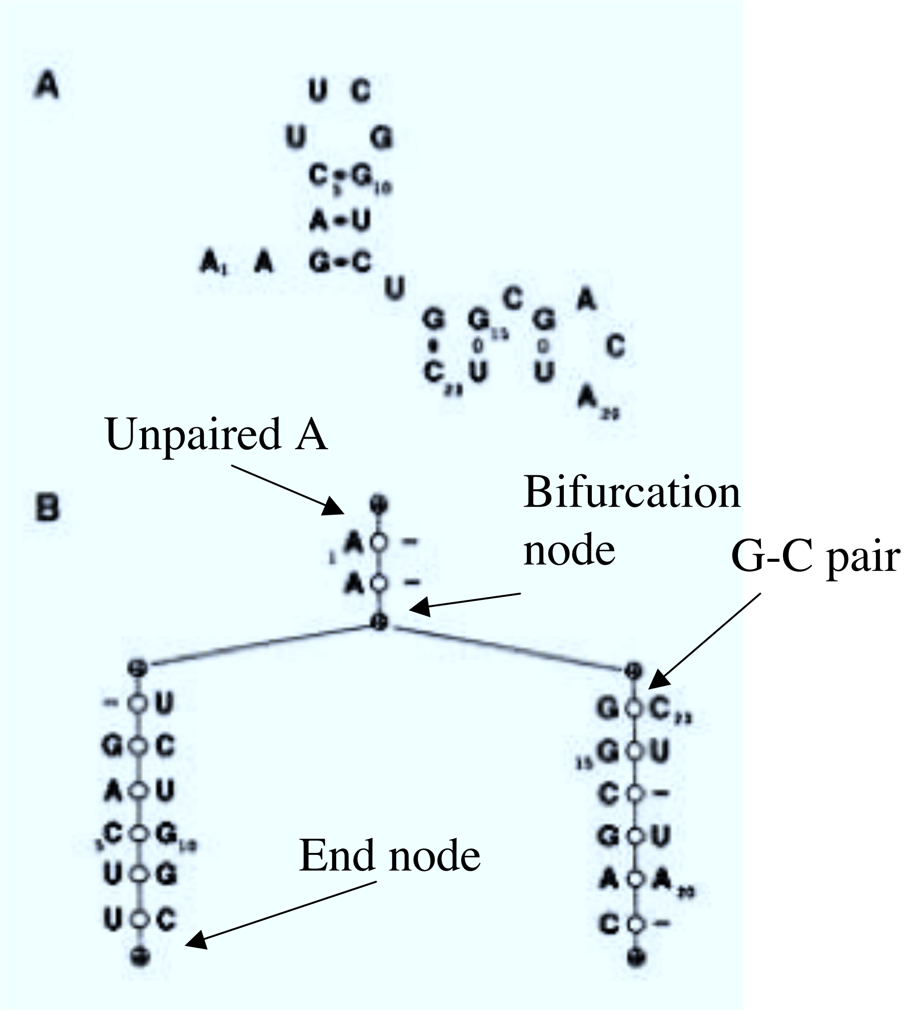
Eddy et al., *RNA sequence analysis using covariance models*, 1994

- CMs based on “guide tree:”
  - Binary tree where nodes correspond to columns in input multiple alignment
  - Models consensus structure of RNA family

CM guide trees

Consensus structure of RNA family

Guide tree
- equivalent to parse tree of a SCFG!
- nodes = paired-bases



Unpaired A

Bifurcation node

G-C pair

End node

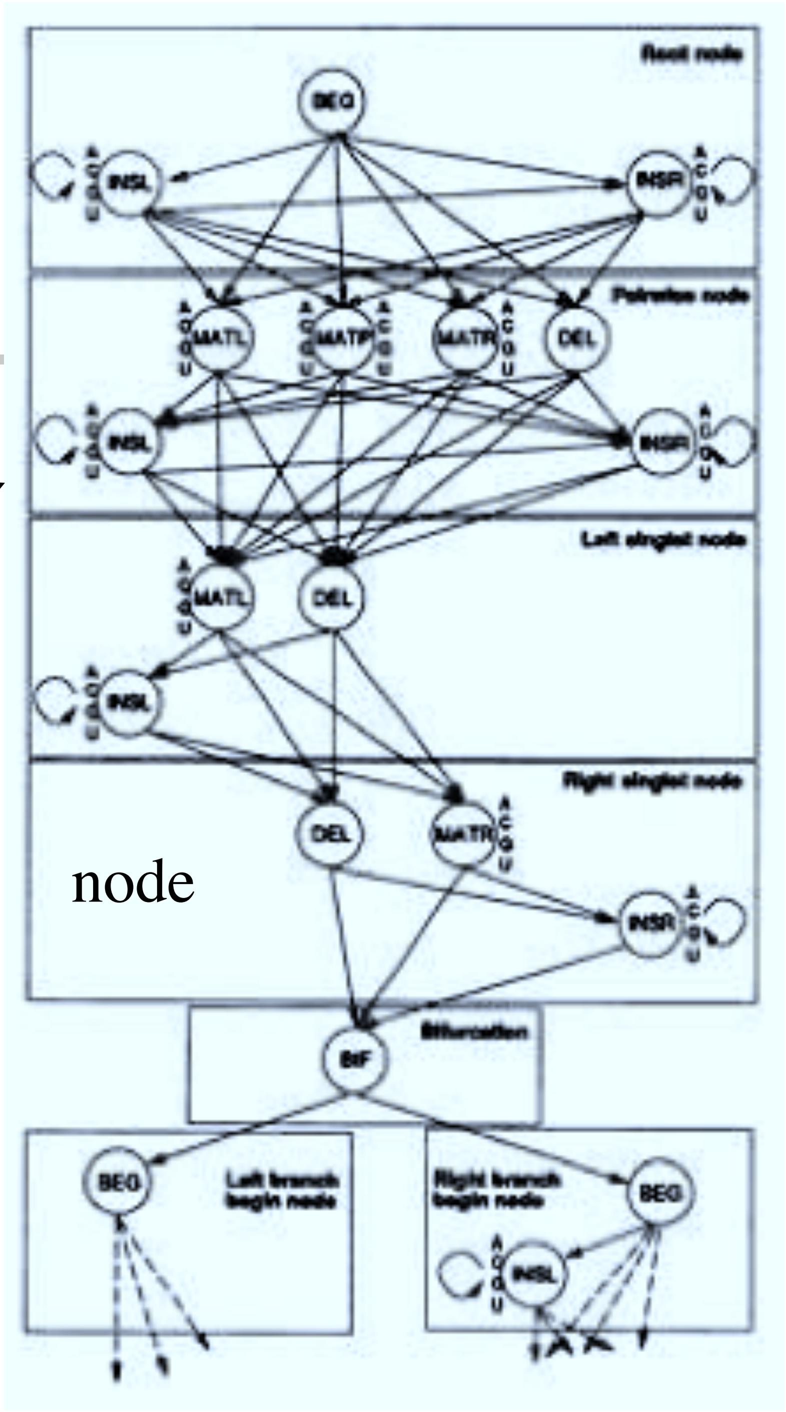
CM intuition

- Model variations in emitted bases
  - Nodes emit bases (pairs) probabilistically

- Model variations in structure
  - Nodes replaced with state machines
  - States for emitting pairs, unmatched pairs, inserts, deletions, etc.
  - States connected via transition probabilities

CM example

- Nodes expanded to state machines
- Ex: Pairwise node
  - Many states
    - MATP - emit a matched base-pair
    - MATR - emit right base of a base-pair
    - INSR - insert unmatched right base
    - DEL - emit nothing, thus delete a base
    - …



CM algorithms

- Align RNA sequence to CM, calculate alignment score
  - **Inside algorithm** for CMs
  - Key difference: uses “Viterbi assumption”
    - prob[CM emits sequence] ~= Prob[Viterbi alignment]
  - Basis for all other CM algorithms
- Search database for homologous RNAs
  - Score subsequences using Inside

CM algorithms

- CM Training - find CM that maximizes likelihood of generating training seqs
  - Given initial alignment
  - Estimate CM structure using “mutual information”
    - How correlated are 2 columns in the alignment?
    - DP algorithm finds tree with consensus secondary structure that maximizes correlation information
  - Use EM to optimize CM’s parameters
    - Align each training sequence to CM
  - Re-estimate new CM structure
  - Repeat until convergence

Paper’s results

- Construct 3 CMs from 1415 aligned tRNAs
- Use CMs to create alignments for test set of sequences
  - 93% correct alignments
  - 90-92% correct using unaligned training seqs!
- Database search compared to TRNASCAN
  - 99.8% true positives, <0.2 false positives/Megabase
- Tertiary structure information adds only ~2-3 bits of correlation information
  - Tertiary info not crucial for database searching?

Discussion

- CMs are alternate formalism of SCFGs
  - But allow for insertions, deletions relative to consensus
  - SCFGs - ungapped models, CMs - gapped models

- CMs are to SCFGs as profile-HMMs are to match-state-only HMMs

Taking phylogeny into account

Knudsen and Hein, *RNA secondary structure prediction using stochastic context-free grammars and evolutionary history*, 1999

Knudsen and Hein, *Pfold: RNA secondary structure prediction using stochastic context-free grammars*, 2003

- Idea: combine info from phylogenetic tree of sequences into SCFGs to improve secondary structure prediction

SCFGs + phylogenetic trees

- Goal: given RNA seqs structural alignment + phylogenetic tree, produce consensus secondary structure

- 2 part model from initial alignment:
  - SCFG - Inside-Outside training
  - Mutational/evolutionary model
    - Matrices of estimated mutation rates between all bases X and Y and pairs XY and X'Y'

Algorithms

- Prob[Alignment | Tree, Model]
  - Needs column probs in alignment
    - Calculated from mutation rates + tree
  - Extend view of grammar as generating columns in the alignment
  - Apply CYK algorithm to new grammar
- ML estimate of tree if not given
  - Assumes input tree topology

Paper's results

- Build KH-99 model from tRNA and LSU rRNA database
  - Mutation rates estimated from counts in database alignment
  - SCFG parameters estimated using Inside-Outside
- Apply model to predict structure of 4 bacterial Pnase P RNA sequences
  - Accuracy improves with # of sequences
  - Phylogenetic info adds ~5% accuracy
- Compared results to CMs
  - Comparable results using less input sequences

Pfold

- Improvements to previous method
  - faster
  - More robust to initial alignment errors
  - Tree estimation faster (scraps ML)
  - Use alternative algorithm to CYK

- Results
  - Pfold implementation - still used today!
  - Similar results, but faster
  - More evolutionary distance yields better accuracy

Detecting noncoding RNAs (ncRNAs)

Rivas and Eddy, *Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs*, 2000

Rivas and Eddy, *Noncoding RNA gene detection using comparative sequence analysis*, 2001

- ncRNA genes contain less statistical signal than protein-coding genes
- How do probabilistic methods function with this weak signal?

Methods and results

- Try #1 - scan genome using SCFG model
  - Detection b/c of C-G composition bias, *not* b/c of structural signal

- Try #2 - scan pairwise alignment of genomes using Pair-SCFG model
  - identify regions with patterns of mutations that suggest a conserved secondary structure
  - Problem: need structurally aware initial alignment
  - Soln: re-align genomes to model…too slow!

Speeding up Pair-SCFG algorithms

Holmes and Rubin, *Pairwise RNA structure comparison with stochastic context-free grammars*, 2002

- Speed up CYK and Inside for Pair-SCFGs
  - Assumes guess at secondary structure of alignment
  - Constrain DP algorithms to only consider pairs of subsequences consistent with structure
  - Calculates “fold envelopes” - set of OK subsequences

- In best case, can lead to linear time CYK and Inside implementations!

SCFG design considerations

Dowell and Eddy, *Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction*, 2004

- Develops a number of small SCFGs and analyzes their prediction accuracy
  - Tradeoff between grammar size and accuracy
  - Knudsen and Hein’s Pfold grammar performs best!
- One-to-one correspondence between sequences and parse trees key to proper functioning of CYK algorithm
  - “structural ambiguity”

Non-probabilistic methods

- Minimum Free Energy (MFE) methods
  - Best structure minimizes free energy of all bonds
  - Mfold and RNAfold
  - Many techniques for incorporating comparative sequence analysis
  - “gold-standard” for RNA secondary structure prediction

- Maximum Weighted Matchings
  - Graph: vertices are bases in sequence, edges with weights from thermodynamic info
  - Max weight matching <=> secondary structure
  - Can predict tertiary interactions!

Summary

- Looked at original papers on SCFG-based and CM-based RNA analysis methods
- Extensions to SCFG models to consider phylogenetic information
- Considered harder problem of detecting ncRNAs
- Briefly looked at SCFG design considerations
- Overview of non-probabilistic methods