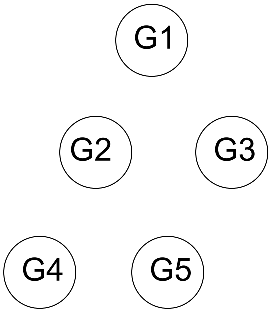
"Change is the end result of all true learning."
-- Leo Buscaglia

Independent Events



If G1, …, G5 are independent, then the joint probability
p(G1, G2, G3, G4, G5)
= *p*(G1) *p*(G2) *p*(G3) *p*(G4) *p*(G5)

Example:

D="Dowell gives the lecture today".
R="It is raining outside today"
Whether it is rain or shine outside doesn't affect whether Dowell is giving the lecture today.
*P(D,R) = p(D) * p(R)*

Conditional Probability Distributions

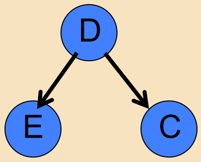
- Conditional probability distributions: $p(B|A)$ = the probability of *B* given *A*.

Example:
D="Dowell gives the lecture today".
E="today' s lecture contains equations"
P(E, D) = Probability that Dowell gives the lecture today and today' s lecture contains equations = 0.05.
P(D)=1/10 = 0.1.

P(E|D) = P(E, D) / P(D) = 0.05/0.1 = 0.5.

Conditional Independence

Example:
D="Dowell gives the lecture today".
E="today's lecture contains equations"
C="today's slides are in Comic Sans font"

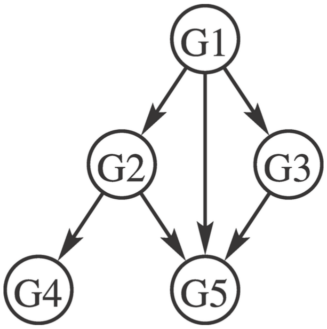


If Dowell is giving the lecture today, then whether today's lecture contains equations doesn't affect whether today's slides are in Comic Sans.
P(E|D,C) = P(E|D)
E and C are conditionally independent given K.

- In Bayesian networks, each node is independent of its non-descendants, given its parents in the graph.
- Using conditional independence between variables, the joint probability distribution of the models may be represented in a compact manner.

Can capture biological relationships

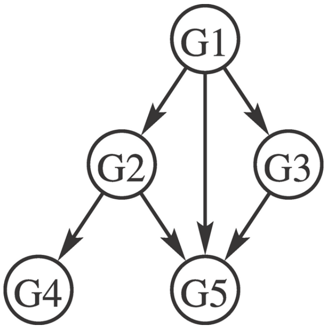
- A directed acyclic graph (DAG) such that the nodes represent mRNA expression levels and the edges represent the probability of observing an expression value given the values of the parent nodes.
- The probability distribution for a gene depends only on its regulators (parents) in the network.



Example: G4 and G5 share a common regulator G2, i.e., they are conditionally independent given G2.
→ factorization of the full joint probability distribution into component conditional distributions.

Needham et al. PLOS Comp Bio 2007

Joint Probability Distribution

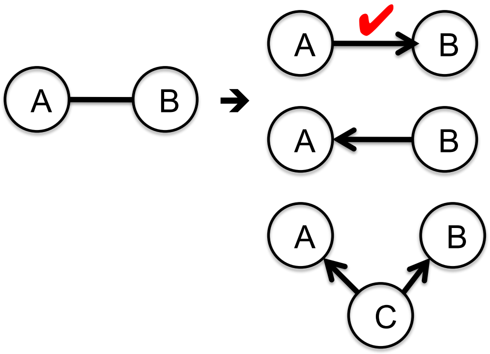


p(G1, G2, G3, G4, G5)
= *p*(G1) *p*(G2|G1) *p*(G3|G1) *p*(G4|G2) *p*(G5|G1, G2, G3)

What kinds of data contain potential information about gene networks?

Large expression sets

- Co-expression (correlation of expression levels) implies connectivity
- But correlation ≠ causality



Adding causality

- Genetic perturbation: DNA variation at A influences RNA variation at B.
- Time series: A goes up prior to B.
- Prior knowledge

Bayesian networks

- Advantages:
  - Compact and intuitive representation
  - Integration of prior knowledge
  - Probabilistic framework for data integration
- Limitation: no feedback loop → dynamic Bayesian networks (variables are indexed by time and replicated in the network)
- References:
  - Using Bayesian Network to Analyze Expression Data. Friedman et al. *J. Computational Biology* 7:601-620, 2000.
  - A Primer on Learning in Bayesian Networks for Computational Biology. Needham et al. PLOS Computational Biology 2007.

Using Bayesian Networks

- There are algorithms for inferring Bayesian networks from large collections of data.
- Given a particular network, can determine whether particular datasets are consistent with the inherent probabilistic relationships implied by the graph.
- Ultimately powerful for predicting the impact of a perturbation.

"We have only just hit the tip of the iceberg.
There's a whole world of this noncoding RNA. "

-- Peter Schultz (2005)

Science
17 May 2002
Vol. 296 · No. 5571
Pages 1189–1352 $9
RNA Silencing and Noncoding RNA
American Association for the Advancement of Science

What do they do?

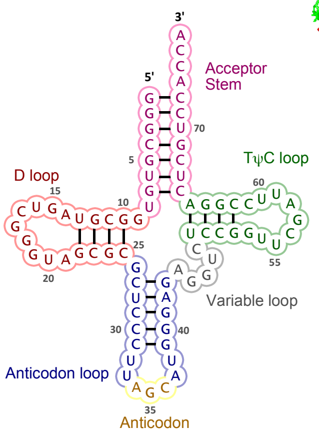
- RNA-protein machine:
  - Transfer RNA (tRNA).
  - Ribosomal RNA (rRNA).
  - RNAs (snRNAs) in spliceosome.
- Catalytic RNAs (ribozymes): catalyzing some functions.
  - 1989 Nobel Prize in Chemistry
- Micro RNAs (miRNAs): regulatory roles.
- Small interfering RNAs (siRNAs): RNA silencing
  - The genome’ s immune system. [Plasterk, *Science* (2002)]
  - 2006 Nobel Prize in Medicine

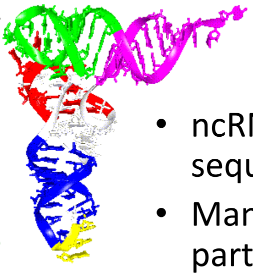
What do they do?

- Riboswitch RNAs: a genetic control element, to control gene expression.
  - found in bacteria, archea, and plants.
- Small nucleolar RNAs (snoRNAs): help the modification of rRNAs.
- tmRNA (tRNA like mRNA): direct abnormal protein degradation.

- lncRNAs: long noncoding RNAs (look like protein coding in terms of exon/intron structure, histone marks, etc)
- eRNAs: enhancer regulation/function?, unstable

RNA secondary structure



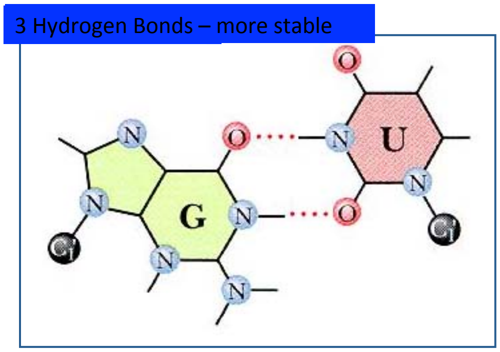


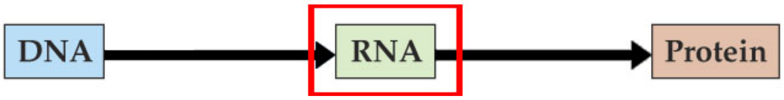
- ncRNA is not a random sequence
- Many RNAs fold into particular base-paired secondary structure (intra-molecular).

RNA secondary structure -- a set of non-crossing base pairs

Many ncRNAs conserve structure

- RNA bases A,C,G,U
- Canonical Base Pairs
  - A-U
  - G-C
  - G-U
    "wobble" pairing
  - Bases can only pair with **one** other base.



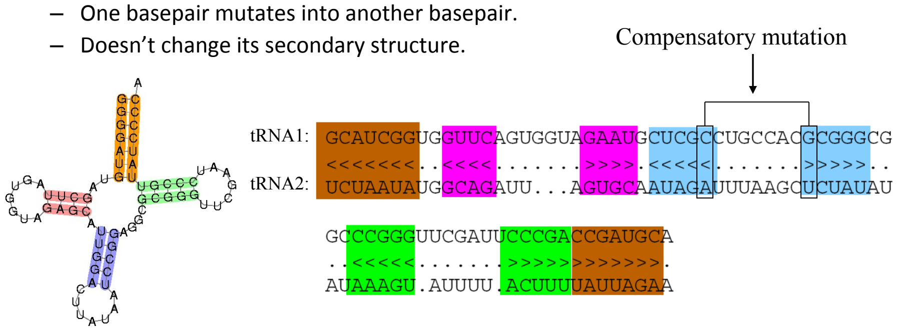


Note: all 16 (including non-canonical) base pairs are actually possible and occasionally observed!!

Image: http://www.bioalgorithms.info/

ncRNA evolution is constrained by it secondary structure

- Drastic sequence changes can be tolerated, so long as structure conserved.
- *Compensatory mutations* are very common.
  - One basepair mutates into another basepair.
  - Doesn't change its secondary structure.



Compensatory mutation

```
tRNA1: GCAUCGGUGGUUCAGUGGUAGAAUGCUCGCCUGCCACGCGGGCG
       <<<<<<<..<<<<.......>>>>.<<<<<.......>>>>>..
tRNA2: UCUAAUAUGGCAGAUU...AGUGCAAUAGAUUUAAGCUCUAUAU

       GCCCGGGUUCGAUUCCCGACCGAUGCA
       ..<<<<<.......>>>>>>>>>>>>.
       AUAAAGU.AUUUU.ACUUUUAUUAGAA
```

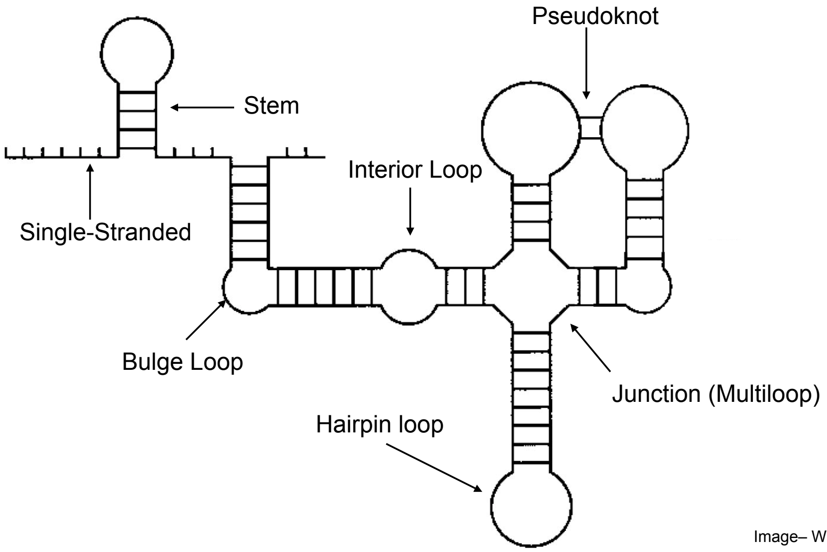
http://www.sanger.ac.uk/Software/Rfam/

Hence sequence alignment also often "fails" to find ncRNAs!!!

RNA secondary structure prediction

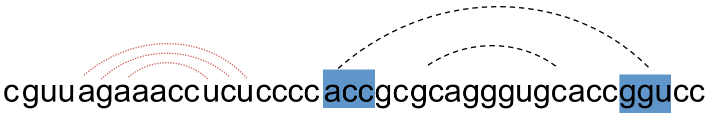
- It is a basic issue in structural ncRNA analysis
- It is important information towards function
- Searching and alignment algorithms are based on these models
- But it is computationally VERY expensive.

RNA secondary structure elements



Image– Wuchty

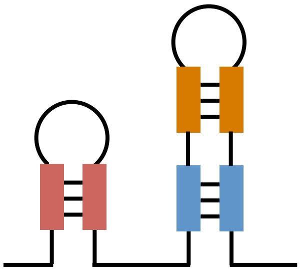
Stems in nested or parallel pattern



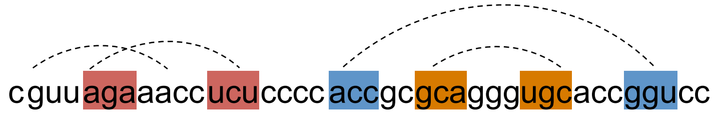
cguuagaaaccucuccccaccgcgcagggugcaccggucc

stem (double helix):
stacked base pairs

loop: strand of
unpaired bases

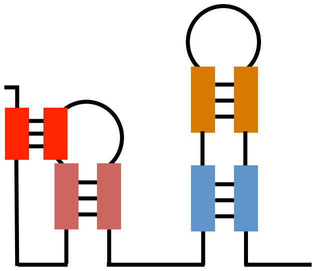


Stems in crossing patterns are pseudoknots



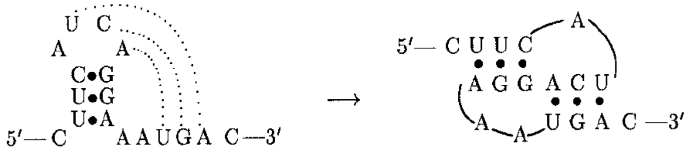
cguuagaaaccucuccccaccgcgcagggugcaccggucc

Pseudoknots: crossing patterns of stems.



Pseudoknots

- Pseudoknots are important for certain ncRNAs
- Violate the non-crossing assumption.
- Pseudoknots make most problems a LOT harder
- We assume there are no pseudoknots, unless otherwise noted.

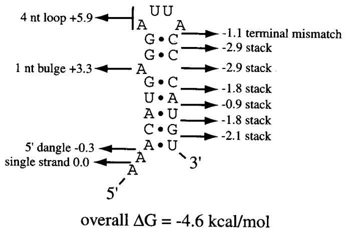


[Rivas and Eddy (1999)]

RNA secondary structure prediction

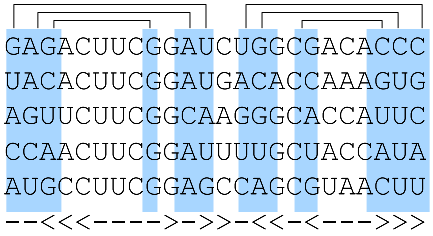
ab inito structure prediction

to predict the structure of a single sequence

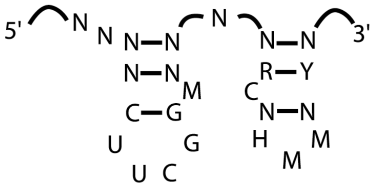


Consensus structure prediction

to predict the structure shared by more than one sequences

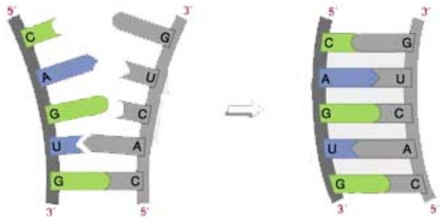


```
GAGACUUCGGAUCUGGCGACACCC
UACACUUCGGAUGACACCAAAGUG
AGUUCUUCGGCAAGGGCACCAUUC
CCAACUUCGGAUUUUGCUACCAUA
AUGCCUUCGGAGCCAGCGUAACUU
--<<<---->->>-<<-<--->>>
```



Specialized intra-molecular "sequence alignment" as a method to determine structure

- Bases pair in order to form backbones and determine the secondary structure
- Aligning bases based on their ability to pair with each other gives an algorithmic approach to determining the optimal structure

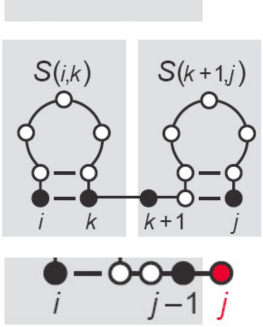


Base Pair Maximization – Dynamic Programming Algorithm

S(i,j) is the folding of the subsequence of the RNA strand from index i to index j which results in the highest number of base pairs

Maximizing Base Pairir

$$S(i,j) = \max \begin{cases} S(i+1,j-1)+1 \quad \text{[if } i,j \text{ base pair]} \\ S(i+1,j) \\ S(i,j-1) \\ \max_{i<k<j} S(i,k)+S(k+1,j) \end{cases}$$



Bifurcation

Base pair at i and j

Unmatched at i

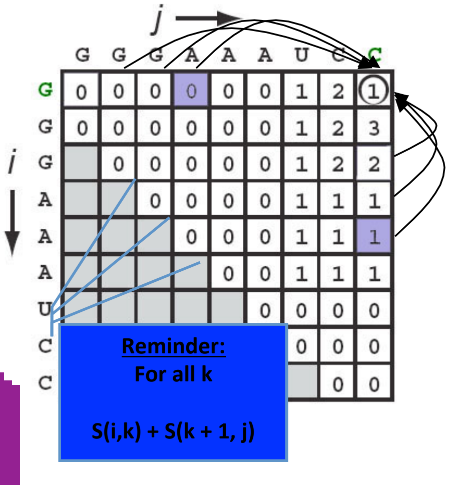
Umatched at j

Images – Sean Eddy

Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Bifurcation – add values for all k



	G	G	G	A	A	A	U	C	C
G	0	0	0	0	0	0	1	2	1
G	0	0	0	0	0	0	1	2	3
G		0	0	0	0	0	1	2	2
A			0	0	0	0	1	1	1
A				0	0	0	1	1	1
A					0	0	1	1	1
U						0	0	0	0
C							0	0	0
C								0	0

Reminder:
For all k

S(i,k) + S(k + 1, j)

Images – Sean Eddy

Dynamic Programming Algorithm

$$S(i,j) = \max \begin{cases} S(i+1,j-1) + \boxed{\mathrm{M}(i,j)} \\ S(i+1,j) \\ \mathrm{S}(i,j-1) \\ \max_{i<k<j} S(i,k) + S(k+1,j) \end{cases}$$

Note that this is the basic folding algorithm, but we can CHANGE the scoring scheme!

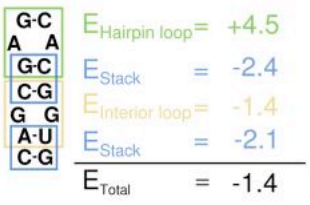
Base Pair Maximization - Drawbacks

- Base pair maximization will not necessarily lead to the most stable structure
  - May create structure with many interior loops or hairpins which are energetically unfavorable
- Comparable to aligning sequences with scattered matches – not biologically reasonable

Alternative scoring scheme:
Energy Minimization
(Thermodynamic Stability)

Energy Minimization Drawbacks

- Compute only one optimal structure
- Usual drawbacks of purely mathematical approaches
  - Similar difficulties in other algorithms
    - Protein structure
    - Exon finding



Alternative scoring scheme:
Probabilistic Framework

Probabilistic Models

- S(i,j) = Score at indices i and j in RNA

$$S(i,j) = \max \begin{cases} S(i+1, j-1) + M(i,j) \\ S(i+1, j) \\ S(i, j-1) \\ \max_{i<k<j} S(i,k) + S(k+1, j) \end{cases}$$

$$M_{i,j} = \sum_{x_i, x_j} f_{x_i x_j} \log_2 \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}}$$

Frequency of seeing the symbols (A, C, G, T) together in locations i and j depending on symbol.

Independent frequency of seeing the symbols (A, C, G, T) in locations i or j depending on symbol.

- Frequencies obtained by aligning model to “training data” – consists of sample sequences
  - Reflect values which optimize alignment of sequences to model

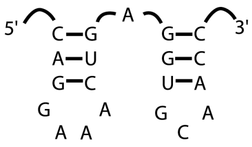
Stochastic context-free grammars

- Big brother to HMMs
  - Emission and transition probabilities!
- Still folding styled dynamic programming

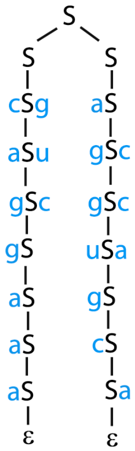
S → aSa' | aS | Sa | SS | ε

SCFGs describe structure with “rules” (aka states).

S → aSa' | aS | Sa | SS | ε

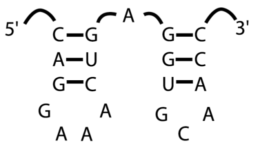


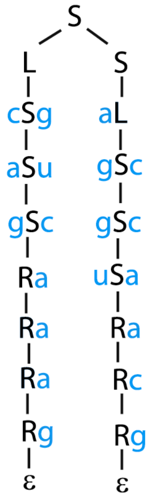
S → SS → cSgS → caSugS → cagScugS →
caggScugS …



Representational ambiguity

S → aSa' | aL | Ra | LS
L → aSa' | aL
R → Ra | ε





Grammar designs vary and each capture different "features" of structure

G1: S → aSa' | aS | Sa | SS | ε

G2: S → aSa' | aL | Ra | LS
L → aSa' | aL
R → Ra | ε

G3: S → aS | T | ε
T → Ta | aSa' | TaSa'

G5: S → L | LS
L → aFa' | a
F → aFa' | LS

G4: S → aSa'S | aS | ε

Standard Algorithms

(notice parallels to HMMs!)

- Scoring (probability of parse tree)

$$P(x,\pi \mid G,\Theta)$$

- Viterbi / CYK (highest probability parse tree)

$$\operatorname{argmax}_{\pi} P(x,\pi \mid G,\Theta)$$

- Inside (probability of sequence, akin to “Forward”)

$$P(x \mid G,\Theta) = \sum_{\pi} P(x,\pi \mid G,\Theta)$$

ncRNA gene finding (a computational challenge)

de novo prediction

identify ncRNA regions from genomic sequence

Profile based methods

identify new instances of a known family of structural ncRNAs

Classification Methods

use evolutionary signatures to distinguish coding, noncoding, and other regions