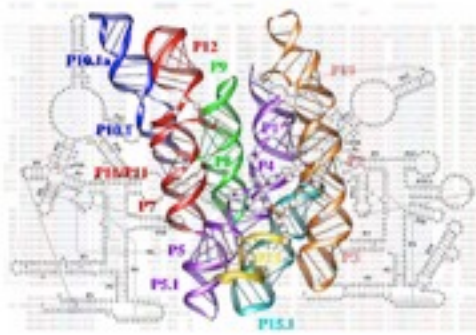
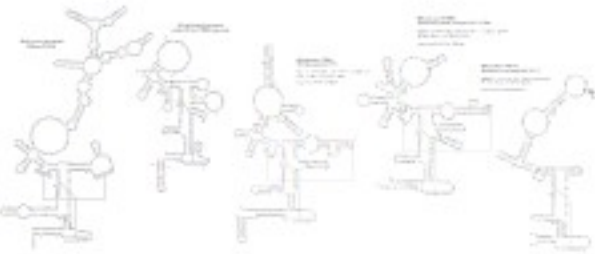


## Comparative Analysis of RNA structure



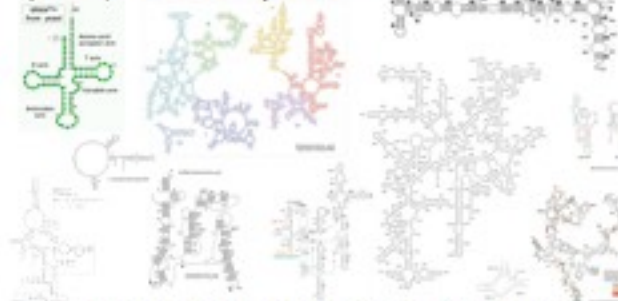
## Comparative Analysis of RNA structure

The process of extracting structural information about a type of RNA from the similarities and differences between different examples of that RNA.



## Comparative Analysis of RNA structure

Comparative analysis is the way to DETERMINE (not predict) RNA secondary structure



All familiar RNA secondary structures were determined by comparative analysis

## Stages of a comparative analysis

- Initial definition of the basic secondary structure
  - A few sequences
  - May use thermodynamic prediction or probing data
  - Resolution: helices
- Refinement of 2<sup>nd</sup> structure & identification of tertiary contacts
  - >50 sequences
  - Identify base-base interactions and higher-order covariations
  - Resolution: base-pairs
- Tertiary modeling
  - local structures
  - assembled global structure
  - Resolution: angles and distances



## Initial definition of 2° structure

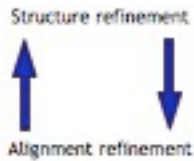
CSA of secondary structure is based on the observation that secondary (and higher-order) structure is conserved despite sequence variation

P4 and P6 in bacterial RNase P RNA



## Initial definition of 2° structure

Comparative analysis is an iterative process: newly-identified structure allows the alignment to be refined so that new structure can be identified.



Increasingly disparate sequences are added when it becomes possible, or when the number of sequences allows, subsets can be analyzed separately to examine structure unique to them

## Initial definition of 2° structure

Initial secondary structure of the *E. coli* RNase P RNA

Each HELIX is "proven" by the presence of covariation in 2 or more basepairs.



## Refinement of 2° & ID of tertiary contacts

As the number of sequences increases, the basic secondary structure can be refined to high resolution using statistical methods to identify and quantitate sequence covariation

$$H = -\sum_b f_b \ln f_b$$

$H$  is a measure of the variability of a sequence position

$b$  in set {A G C U} or {A•A A•G A•C A•U G•A G•G G•C ... U•U}

$f_b$  = frequency of each base or base pair

$$M(x,y) = H(x) + H(y) - H(x,y)$$

$x$  and  $y$  are pairs of positions in an alignment

$$M(x,y) \leq \min[H(x), H(y)]$$

$$\chi^2 = 2(M \times n)$$

$M(x,y)$  is maximized when both positions are highly variable and also perfectly correlated

## Refinement of 2° & ID of tertiary contacts

### Example M(x,y) analysis of a basepair

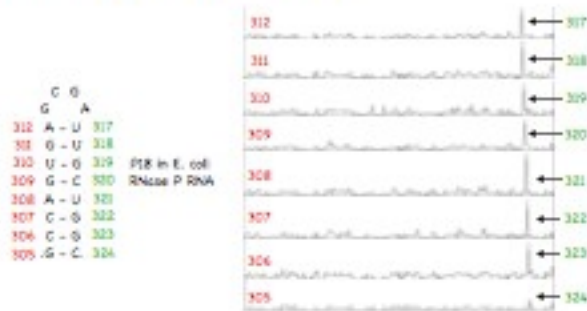
- Position 234 of bacterial RNase P against all other positions in the alignment



Position 234 of bacterial RNase P RNA basepairs with position 247

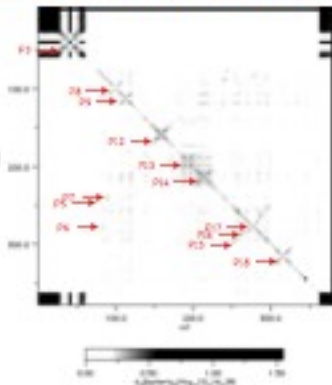
## Refinement of 2° & ID of tertiary contacts

### Example M(x,y) analysis of a helix



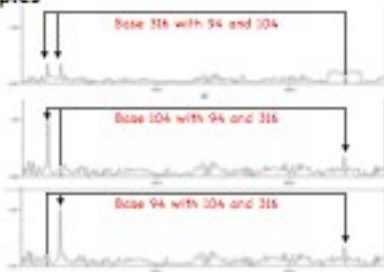
## Refinement of 2° & ID of tertiary contacts

### 2-dimensional 'Dotplot' of M(x,y) of bacterial RNase P RNA



## Refinement of 2° & ID of tertiary contacts

### Base triples



These are identified in the same way as secondary basepairs, except that all three bases covary with each other



## Refinement of 2° & ID of tertiary contacts

Comparative analysis is the gold standard for the determination of RNA secondary structures.

Some RNA secondary structures determined by CSA:

tRNA	splicing snRNAs	T-box RNA
ssu rRNA	guide snoRNAs	Riboswitches
lsu rRNA	6S RNA	Txn terminators
5S rRNA	tmRNA	Attenuators
5.8S rRNA	RNase MRP RNA	Rep origin regulators
RNase P RNA	telomerase RNA	antisense RNAs
SRP RNA	hammerhead ribozyme	Editing guide RNAs
Group I introns	hairpin ribozyme	SELEX aptamers!
Group II introns	delta virus ribozymes	&c, &c ...

## Refinement of 2° & ID of tertiary contacts

Comparative analysis is the gold standard for determination of RNA secondary structures, but...

### Strengths:

- objective, quantitative
- automatable & visualizable
- basepair resolution
- can distinguish thermodynamically equivalent possibilities
- only biologically-relevant structures identified
- identifies any base-base interaction with alternative versions

### Weaknesses:

- phylogenetic effects add complexity
- sequence sample effects
- alignment basically a manual process
- May best for final stages of secondary analysis & analysis of tertiaries
- no specific information from invariant sequences
- no specific information from idiosyncratic sequences
- difficult to incorporate biochemical data

## Tertiary modeling

### Independent substructures

- hairpin elements
- large insertions/deletions

### Helical stacks

- covariation in length
- steric constraint/adjacency

### Secondary structure motifs

- replaceable alternatives
- structure classes

### Phylogenetic minimum core

- essential vs stabilization vs trivial
- inside vs outside

### Assembly of global models



## Tertiary modeling

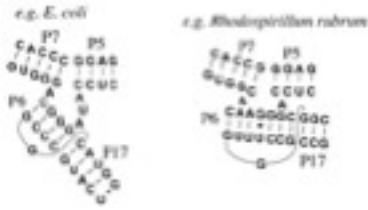
### Helical stacking



Stacking is a major force in directing the folding of RNA

## Tertiary modeling

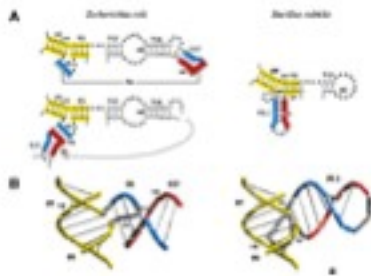
### Helical stacking



Helices in pseudoknots usually stack

## Tertiary modeling

### Local structure models



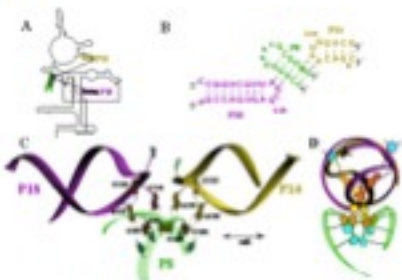
## Tertiary modeling

### Another GNRA:receptor interaction



## Tertiary modeling

### Base triples



This is a common GNRA:minor groove tertiary interaction

## Tertiary modeling

### Phylogenetic minimal core structure



The phylogenetic minimum core contains all of the essential sequence and structure. Other elements generally contribute (importantly or not) to stability, or are trivial elements that are tolerated as long as they do not interfere with the rest of the molecule.

---

---

---

---

---

---

---

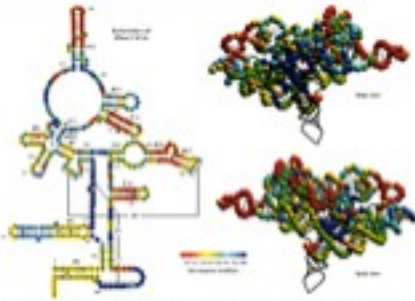
---

---

---

## Tertiary modeling

### Conserved inside, variable outside



Variable sequences and helices of conserved length are concentrated in the heart of the RNA. Variable sequences and structures are peripheral.

---

---

---

---

---

---

---

---

---

---

## Tertiary modeling

But just because you know the structure of all the parts doesn't mean you know the structure of the whole!



---

---

---

---

---

---

---

---

---

---

## Tertiary modeling

### The need for constraints

- The number of possible foldings of a polymer of any length are astronomical
- Constraints are bits of information that limit the possibilities, i.e. they constrain the model.
- Some constraints:
  - The secondary structure!
  - Phylogenetic variation
  - Tertiary interactions
  - Helical stacking data
  - Distance measurements & crosslinks
  - Known overall shape
  - etc, etc...

---

---

---

---

---

---

---

---

---

---

## Tertiary modeling

Interactive computer modeling : Massire & Westhof



Constraint #1 : the detailed secondary structures of type A and B RNase P RNAs

---

---

---

---

---

---

---

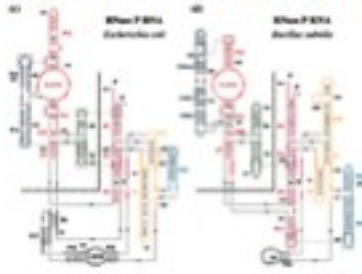
---

---

---

## Tertiary modeling

Interactive computer modeling : Massire & Westhof



Constraints #2 & 3: All known tertiary interactions & stacking partners (including some suggested by the preliminary models, and some reasonable guesses)

---

---

---

---

---

---

---

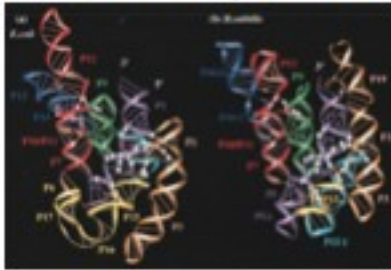
---

---

---

## Tertiary modeling

Interactive computer modeling : Massire & Westhof



Pieced all together to satisfy crosslinking data, biochemical data, and aesthetics, and...Violal

---

---

---

---

---

---

---

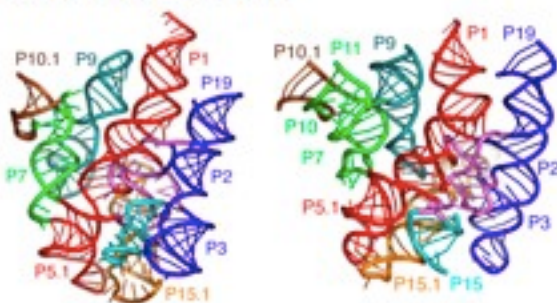
---

---

---

## Tertiary modeling

Models vs crystal structure



Comparative model

Crystal structure

---

---

---

---

---

---

---

---

---

---

## Stages of a comparative analysis

- Initial definition of the basic secondary structure
  - A few sequences
  - May use thermodynamic prediction or probing data
  - Resolution: helices
- Refinement of 2<sup>o</sup> structure & identification of tertiary contacts
  - ~50 sequences
  - Identify helical basepair interactions and higher-order covariations
  - Resolution: base-pairs
- Tertiary modeling
  - local structures
  - assembled global structure
  - Resolution: angles and distances

---

---

---

---

---

---

---

---

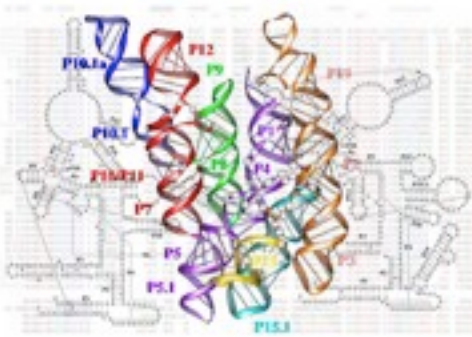
---

---

---

---

## Comparative Analysis of RNA structure



---

---

---

---

---

---

---

---

---

---

---

---