**Specific Aims**

**Title: Phylogenetic tests for conservation of multiple structures in families of noncoding RNAs.**

**Background:**

One substantial product of ongoing multi-decade efforts to re-evaluate the central dogma of molecular biology is available to biologists in the form of huge databases of untranslated RNAs (ncRNAs) such as Rfam. Grouped into families sharing similar secondary (and presumably 3d) structures, conserved ncRNAs have been demonstrated to play roles in a wide variety of processes in all domains of life.

In the cases of some ncRNA families, the specific role of a structure in the biological activity of RNAs have been deduced by biochemistry and can be understood in enzymatic terms much like proteins. In other cases, although candidate structures have been deduced by methods ranging from in-silico modeling (via thermodynamic or phylogenetic models) to in-vitro imaging (such as NMR), the specific role of structure in function remains mysterious. Moreover, although most ncRNAs can fold into multiple secondary structures at various negative thermodynamic potentials (suboptimal structures), most descriptions of RNA families are phrased in terms of single ‘consensus’ structures.

Noting that some well-characterized ncRNAs assume various conformations in their biological roles (e.g.: riboswitches, self splicing introns), this project constitutes a systematic investigation of the likelihood that suboptimal structures are conserved in a subset of the ~fourteen thousand Rfam ncRNA families by a novel phylogenetic method.

**Aims and Innovations:**

**1) Phylogeny and structure aware enumeration and classification of mutation events in ncRNA families via max-likelihood ancestor reconstruction.**

State of the art algorithms use a combination of thermodynamic modeling and phylogenetic analysis to deduce single predictions for optimal foldings of ncRNA families. One such algorithm is implemented by the program RNAalifold in the widely used Vienna RNA package. RNAalifold augments a straightforward thermodynamic computation of structure for a familial consensus sequence by adding negative pseudo-energies to base positions where compensated mutations occur in the sequence alignment for a family.

By counting compensated mutation observed in raw sequence alignments however, RNAalifold and similar programs fail to compute the true frequency of compensated mutation that would be observed in a parsimonious phylogenetic tree for a given ncRNA family. Using PAML for inference of maximum likelihood ancestor sequences in a family, this project supposes that it derive more evolutionarily accurate evidence for conservation of base pairing in ncRNA secondary structures.

**2) Prediction of conservation profiles for ncRNA family (sub)optimal structures.**

After establishing a set of candidate functional structures with an out-of-the-box algorithm and evaluating phylogenetic evidence for conservation of each structure as described in the previous step, this project will compute expected conservation signals for evolution acting on each ncRNA structure. To lowest order, this signal would amount to observation of (1) exclusively base pair-preserving mutations at paired sites (2) randomly distributed mutations at unpaired sites.

Noting that not all sites in the primary sequence and secondary structure of an ncRNA are equally important in determining folded conformation of ncRNA, this project will implement a more principled prediction for sequence/pairing conservation using an “importance” metric derived from the program RNAmutants. RNAmutants tests RNA structure stability over all possible k-fold mutations. To determine the importance of base-pairing of sequence elements (i,j) in structure *S*, I will compute the thermodynamic partition function of constrained folding to structure *S* over all 2-fold mutations of bases (i,j)*.* From an importance metric thus computed over sequences in a family, I will predict evolutionary signatures of selection acting to preserve elements of the structure ensemble under consideration.

**3) Linear algebraic evaluation of likelihood that conservation is acting to preserve (sub)optimal structure.**

Having **(1)** computed a phylogenetic history of each candidate functional suboptimal folding under consideration for an Rfam family and **(2)** predicted evolutionary signatures for the conservation of each structure, I will deduce a subset of candidate structures bearing responsibility for the observed evolutionary history of an ncRNA family.

To this end, several approaches could be envisioned. The simplest would be to simply rank structures in order of the ratio of compensated (base-pairing preserving) mutations to uncompensated mutations according seen in the family’s evolutionary history. Wishing to directly compare candidate structures and deduce a minimal set of structures explaining the evolutionary history of a family, I propose an alternative approach involving

**(a)**:breaking the ncRNA family into N clusters of similar sequences.

**(b)**:summing evolutionary signals from **(1)** for each of N clusters into N L-dimensional vectors where L is the length of the aligned sequence for the ncRNA family under consideration.

**(c):** using regularized least squares linear regression to express the evolutionary signature vectors from **(b)** in terms of the predicted conservation patterns for each structure from **(2)**.

The structures with highest learned coefficients from regression will be considered the most likely to be under structural selection.