



# NARI'2K

Novel Approaches in RNA Informatics

Book of Abstracts

May 18-19, 2000  
Montréal, Québec

# Index

Invited presentations

**From 20 atoms to 2: Descriptors of RNA conformation.**

Chuck Duarte, Columbia University

**Optimal solution for RNA pseudoknots using dynamic programming: A linguistic interpretation**

Elena Rivas, Washington University

**Computational Paradigms for RNA Structure Analysis and Prediction: A Multifaceted Approach**

Bruce Shapiro, NCI.

**RNA structure based drug design.**

Mohan Venkantraman, Isis Pharmaceuticals Inc.

**Informatic approaches to RNA secondary and tertiary structure prediction**

Chuck Wilson, UCSC.

**Statistics of base pair stack frequencies**

Michael Zuker, Washington University

Abstracts

**Direct visualization of docking of the hairpin ribozyme-substrate complex by atomic force microscopy**

Michael J. Fay, Nils G. Walter and John M. Burke ..... 1

**Dynamics of metal ion binding sites in a catalytic RNA**

Thomas Hermann and Dinshaw J. Patel ..... 2

**Stem Trace: A GUI-based tool for an exploratory RNA structure analysis**

Wojciech Kasprzak and Bruce Shapiro ..... 3

**Modelling the structure of an RNA dimer using Mc-Sym**

Tony Kusalik ..... 4

<b>Quantitative analysis of nitrogen base spatial relations in RNA three-dimensional structures</b>	
Sébastien Lemieux, Patrick Gendron and François Major .....	5
<b>Application of the scan statistic to locate significant palindrome clusters in nucleic acid sequences</b>	
Ming-Ying Leung .....	6
<b>Using computational genetics to identify snoRNAs in Yeast and archaeal genomes</b>	
Todd M. Lowe .....	7
<b>Effect of modified nucleotides and magnesium on <i>Escherichia coli</i> tRNA<sup>Glu</sup> structure and on its aminoacylation by glutamyl-tRNA synthetase</b>	
Eric Madore, Catherine Florentz, Richard Giegé, Shun-ichi Sekine, Shigeyuki Yokoyama and Jacques Lapointe .....	8
<b>Identification of sequence motifs in oligonucleotides whose presence is correlated with antisense activity</b>	
Matveeva OV, Atkins JF, Giddings M, Freier SM, Wyatt JR, Spiridonov AN, Shabalina SA, Gesteland RF and Tsodikov AD .....	9
<b>Use of computer modelling and biochemical data to define intermediates of the folding pathway for the delta ribozyme catalysis</b>	
Martin Pelchat and Jean-Pierre Perreault .....	10
<b>A bend in the hairpin ribozyme substrate binding domain</b>	
Robert Pinard, Dominic Lambert, François Major and John M. Burke .....	11
<b>Basic geometric problems on reconfiguring polymers</b>	
Michael Soss and Godfried T. Toussaint .....	12
<b>Linguistic analysis of the 5S ribosomal RNA of Actinomycetes</b>	
Sandra M.R. Subacius and Gabriel Padilla .....	13
<b>Mining the databases for mobile group II introns</b>	
Steven Zimmerly, Georg Hausner, and Xu-chu Wu .....	15

## Direct visualization of docking of the hairpin ribozyme-substrate complex by atomic force microscopy

Michael J. Fay, Nils G. Walter and John M. Burke

Dept. of Microbiology and Molecular Genetics and the Markey Center for Molecular Genetics,  
University of Vermont, Burlington, VT 05405

The hairpin ribozyme is a fifty nucleotide catalytic RNA, derived from the satellite RNA of tobacco ringspot virus. As a site-specific RNA endonuclease, the 5' and 3' cleavage products have 2'-3' cyclic phosphate and 5' hydroxyl termini, respectively. The reaction mechanism is yet to be fully defined, and the three dimensional structure is unknown. However, it is known that specific interactions between, or docking of, the two internal loop domains of the ribozyme-substrate complex is essential for substrate cleavage.

In recent years, atomic force microscopy (AFM) has been effectively utilized to image biological molecules in aqueous buffers. Tapping mode AFM obtains data by making a raster scan of a surface, recording the height deflection of the probe upon contact with an object immobilized on the surface. The collected data are used to compose a contour plot of the scanned area that may, for example, consist of small nucleic acids bound to atomically flat, freshly cleaved mica. Here, we describe the optimization of imaging buffer and sample preparation for visualization of individual hairpin ribozyme-substrate complexes in solution. With these conditions, we have acquired images which show it is possible to directly observe structural changes between inactive and active variants of this relatively small catalytic RNA molecule.

## Dynamics of metal ion binding sites in a catalytic RNA

Thomas Hermann and Dinshaw J. Patel  
Memorial Sloan-Kettering Cancer Center, New York

RNA three-dimensional structures contain metal cations as intrinsic components which stabilize the folding of the negatively charged sugar-phosphate backbone and participate in the chemistry of RNA function. We have recently developed a prediction method based on Brownian dynamics (BD) simulations of cation diffusion to calculate electronegative pockets in RNA folds [1]. The prediction algorithm exhaustively explores potential metal ion binding sites, reproducing the positions of metal ions in crystal structures and revealing previously unknown cation binding sites in RNA three-dimensional structures. We have used a combination of molecular dynamics (MD) simulations under realistic solvent conditions and BD simulations of cation diffusion in order to study the dynamics of metal ion binding sites in the lead-dependent ribozyme. Solution conformations of the catalytic RNA were obtained by MD simulations starting from the crystal structure of the catalytic RNA [2]. BD simulations performed on succeeding conformers reveal the dynamics of electronegative pockets in the RNA fold. Consequences for metal ion binding and implications on the catalysis of the ribozyme will be discussed.

[1] Hermann and Westhof, *Structure*, 6, 1303 (1998).

[2] Wedekind and McKay, *Nature Structural Biology*, 6, 261 (1999).

## Stem Trace: A GUI-based tool for an exploratory RNA structure analysis

Wojciech Kasprzak<sup>1</sup> and Bruce Shapiro<sup>2</sup>

<sup>1</sup> Intramural Research Support Program, SAIC-Frederick, NCI-FCRDC

<sup>2</sup> Laboratory of Experimental and Computational Biology, NIH, NCI-FCRDC

Stem Trace is an important software tool developed for STRUCTURELAB, an RNA structure analysis computer workbench. It provides the capability to analyze a database of a large number of computationally generated structure conformations in a novel, visually driven, and exploratory way.

Stem Trace is essentially a two-dimensional plot of all unique helical stems from a solution space of structure conformations of a given sequence, presenting an orthogonal view to the usually employed stem histograms (dot-matrices). Displayed information is color-coded to reflect cumulative frequency of occurrence or the stem energy, and it retains an explicit depiction of all the individual conformations used as inputs to it. Besides providing a visual representation of a solution space for an RNA sequence, Stem Trace is also an active graphical user interface (GUI) to the underlying data. A set of functions associated with every trace performs searching, sorting, scaling, thresholding and, through connections with other STRUCTURELAB tools, drawing and labeling of structures based on the data extracted from a specific trace.

The generic nature of such a data presentation paradigm has allowed us to develop Stem Trace into an analysis tool capable of dealing with a range of RNA and DNA folding data for single and multiple sequences, thus introducing some phylogenetic analysis elements. A separate set of functions is devoted to dealing with sequence/structure alignment and structural motif search and analysis.

The tool has proven very useful in the analysis of the solution spaces of the Genetic Algorithm and the Dynamic Programming Algorithm (such as MFOLD) for both RNA and DNA sequences. In its multi-sequence mode it has yielded clear visualization of structure preservation in a set of HIV and SIV strains.

## Modelling the structure of an RNA dimer using Mc-Sym

Tony Kusalik

Department of Computer Science, University of Saskatchewan

Mc-Sym is a system developed at the Université de Montréal for modeling and predicting the 3D structure of RNA. The system is based on a constraint satisfaction algorithm, and allows use of secondary structure and low-resolution data as input. Mc-Sym has successfully generated 3D structures for portions of molecules and whole single molecules. More recently, it has been applied to modeling the structure of multimers. This presentation describes one such effort: modeling the structure of the proposed dimer for inactive E. Coli tRNA-Glu. Inferring the structure of the dimer required a modeling methodology different from that normally used with Mc-Sym. The new methodology follows a cyclic approach where structures are built using pre-screened substructures from a previous cycle. This presentation describes the new methodology, as well as the results obtained for tRNA-Glu to date.

# Quantitative analysis of nitrogen base spatial relations in RNA three-dimensional structures

Sébastien Lemieux, Patrick Gendron and François Major

Université de Montréal, Dept. d'informatique et Recherche Opérationnelle, Montréal, Québec, Canada

**PROBLEM:** The higher-order structure of RNA can be represented by a set of chemical interactions among their nucleotides, such as nitrogen base pairing and s-tacking [1]. Quantitative studies of nitrogen base spatial relations have applications in RNA 3-D motif identification, in the definition of RNA 3-D conformational search spaces, and in the evaluation of the amount of structural information present in the currently available 3-D structures. However, no objective method existed for objectively studying nitrogen base spatial relations.

**METHOD:** We developed a procedure that, given the atomic coordinates of a pair of nitrogen bases, specifically identifies its type of spatial relation. Homogeneous transformation matrices were used to encode spatial relations. The similarity between two spatial relations is assessed using a distance metric that counterbalance the rotation and translation in the context of nitrogen bases. From the distance metric, we defined the peculiarity of a given spatial relation to a set of spatial relations. Finally, we developed a computer program that rapidly analyzes the spatial relations in a RNA 3-D structure, which in particular identifies unusual ones.

**RESULTS:** We extracted, classified and stored in a database the spatial relations of all 3-D structures made available to us. For each type of spatial relation, an ordered list of homogeneous transformation matrices was assigned, which can be used for its conformational sampling. We analyzed the RNA-binding domain of ribosomal protein L11 [2], and found that its most peculiar spatial relations are located in the regions where the RNA directly interacts with the protein. An estimate of the "completeness" of the current 3-D structure database was made.

[1] F. Major et al. (1991) *Science*, 253:1255-1260.

[2] G.L. Conn et al. (1999) *Science*, 284:1171-1174.

## Application of the scan statistic to locate significant palindrome clusters in nucleic acid sequences

Ming-Ying Leung

Division of Mathematics and Statistics, University of Texas at San Antonio

Nonrandom clusters of palindromes have been observed around origins of replication and regulatory sites of some viruses. A method to distinguish statistically significant clusters of palindromes from those occurring by chance would be helpful in locating putative regions of biological interest in viral genomes. Modeling the occurrences of palindromes in nucleic acid sequences as independent and identically distributed random variables uniformly distributed on the unit interval  $(0,1)$ , significant palindrome clusters are located using the  $r$ -scan statistic which has a Poisson type limiting probability distribution. The procedure is illustrated with several complete genomes from the families of herpesviruses and paramyxoviruses.

## Using computational genetics to identify snoRNAs in Yeast and archaeal genomes

Todd M. Lowe

Stanford University School of Medicine, Department of Genetics, Stanford, CA 94305-5120

In eukaryotes, dozens of post-transcriptional modifications are directed to specific nucleotides in ribosomal RNAs (rRNAs) by small nucleolar RNAs (snoRNAs). Many members of this RNA gene family had remained unidentified in *Saccharomyces cerevisiae*, despite the availability of a complete genome sequence. I developed a probabilistic model of the snoRNA gene family and used it to computationally screen the yeast genome [1]. Twenty-two new snoRNA genes were identified and verified experimentally using gene disruptions and other experimental characterization. In all, there are now 41 members of this gene family, and we estimate fewer than four remain unidentified.

SnoRNAs had formerly been identified in eukaryotes exclusively. As part of a collaboration with Patrick Dennis at the University of British Columbia, we have now identified homologs of snoRNA genes in both branches of the Archaea [2]. Eighteen small sno-like RNAs (sRNAs) were cloned from the archaeon *Sulfolobus acidocaldarius* by co-immunoprecipitation with aFIB and aNOP56, the archaeal homologs of eukaryotic snoRNA-associated proteins. We re-trained our probabilistic yeast snoRNA model on the *S. acidocaldarius* genes to search for more sRNAs in archaeal genomic sequences. Over 250 additional sRNAs were identified in eight archaeal genomes representing both the Crenarchaeota and the Euryarchaeota. Small nucleolar RNA based rRNA modification was therefore probably present in the last common ancestor of Archaea and Eukarya, predating the evolution of a morphologically distinct nucleolus.

The complete set of yeast and archaeal snoRNAs can be found on the Eddy lab snoRNAdb website: [rna.wustl.edu/snoRNAdb/](http://rna.wustl.edu/snoRNAdb/).

[1] Lowe and Eddy, *Science*, 283:1168-1171, 19 Feb 1999.

[2] Omer et al., *Science*, 288:517-522, 21 Apr 2000.

## Effect of modified nucleotides and magnesium on *Escherichia coli* tRNAGlu structure and on its aminoacylation by glutamyl-tRNA synthetase

Eric Madore<sup>1</sup>, Catherine Florentz<sup>2</sup>, Richard Giegé<sup>2</sup>, Shun-ichi Sekine<sup>3</sup>, Shigeyuki Yokoyama<sup>3</sup> and Jacques Lapointe<sup>1</sup>

<sup>1</sup> Département de Biochimie et de Microbiologie, Centre de Recherche sur la Fonction, la Structure et l'Ingénierie des Protéines (CREFSIP), Université Laval, Québec, Canada, G1K 7P4.

<sup>2</sup> UPR 9002, Institut de Biologie Moléculaire et Cellulaire du CNRS, 15 rue René Descartes, F-67084 Strasbourg Cedex, France.

<sup>3</sup> Department of Biophysics and Biochemistry, Graduate School of Science, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

*Escherichia coli* tRNAGlu possesses five modified nucleotides: Y13, mnm5s2U34, m2A37, T54 and Y55 and requires magnesium to fold in an active conformation. Over-producing this tRNA in its homologous host results in the presence of several distinctly modified forms of this molecule that we named modivariants. The predominant tRNAGlu modivariant in wild-type *E. coli* contains all five modified nucleosides. Four other overproduced modivariants differ from it by respectively, either the presence of an additional Y, or the presence of s2U34, or the lack of A37 methylation combined with either s2U34 or U34. Chemical probing reveals that the anticodon loop of the predominant modivariant is less reactive to the probes than that of the four others. Furthermore, the modivariant with neither mnm5s2U34 nor m2A37 has additional perturbations in the D- and T-arms and in the variable region. In the absence of magnesium, all these modivariants adopt a new conformation: chemical and enzymatic probing revealed that they have a hairpin structure with two internal loops. The lack of a 2-thio group in nucleoside 34 decreases by 520-fold the specificity of *E. coli* glutamyl-tRNA synthetase for tRNAGlu in the aminoacylation reaction, showing that this thio group is the identity element in the modified wobble nucleotide of *E. coli* tRNAGlu. The modified nucleosides content also influences the recognition of ATP and glutamate by this enzyme, and in this case also, the predominant modivariant is the one that allows the best specificity for these two substrates. These structural and kinetic properties of tRNAGlu modivariants indicate that the modification system of tRNAGlu optimizes the stability of tRNAGlu and its action as cofactor of the glutamyl-tRNA synthetase for the recognition of glutamate and ATP.

## Identification of sequence motifs in oligonucleotides whose presence is correlated with antisense activity

Matveeva OV<sup>1\*</sup>, Atkins JF<sup>1</sup>, Giddings M<sup>1</sup>, Freier SM<sup>3</sup>, Wyatt JR<sup>3</sup>, Spiridonov AN<sup>4</sup>, Shabalina SA<sup>5</sup>, Gesteland RF<sup>1</sup> and Tsodikov AD<sup>2</sup>

<sup>1</sup> Department of Human Genetics, University of Utah, Salt Lake City 84112-5330, USA

<sup>2</sup> Huntsman Cancer Institute, Department of Oncological Sciences, University of Utah, Salt Lake City 84112, USA

<sup>3</sup> ISIS Pharmaceuticals, Carlsbad, California 92008, USA.

<sup>4</sup> IHS, Ithaca, New York 14853, USA

<sup>5</sup> National Center for Biotechnology Information, NLM, NIH, Bethesda, Maryland 20814, USA

\* To whom correspondence should be addressed

Design of antisense oligonucleotide targeting any mRNA can be much more efficient when several activity-enhancing motifs are included and activity-decreasing motifs are avoided. This finding was made after statistical analysis of data collected from more than a thousand experiments with phosphorothioate-modified oligonucleotides. Highly significant positive correlation between the presence of motifs TCCC, ACTC, GCCA, CTCT in the oligonucleotide and its antisense efficiency was demonstrated. In addition to that negative correlation was revealed as well for the motifs GGGG, ACTG, AAA and TAA. It was found that likelihood of activity of oligonucleotide against a desired mRNA target is sequence motif content dependent.

## Use of computer modelling and biochemical data to define intermediates of the folding pathway for the delta ribozyme catalysis

Martin Pelchat and Jean-Pierre Perreault

Département de Biochimie, Université de Sherbrooke, Sherbrooke, Québec, J1H 5N4, Canada

Recently, the structure of hepatitis delta virus (HDV) ribozyme was determined based on crystallographic data using the antigenomic ribozyme and the RNA-binding protein U1A as a crystallization module. However this structure only showed the conformation adopted by the ribozyme after the cleavage step. It has been demonstrated that structural changes are necessary for active complex formation. As our goal is to identify the different structural intermediates of the reaction, we used the molecular modeling software MC-SYM developed by the Major's group as a tool to hypothesize the conformational transitions based on our biochemical data. We were able to define several putative conformational intermediates for this reaction which enable us to illustrate different steps of folding pathway. Moreover, these models could be used to design experiments to access the pathway. This is a new application for the MC-SYM software, as it is an approach to deal with folding pathway of RNA catalysis.

## A bend in the hairpin ribozyme substrate binding domain

Robert Pinard<sup>1</sup>, Dominic Lambert<sup>2</sup>, Francois Major<sup>2</sup> and John M. Burke<sup>1</sup>

<sup>1</sup> University of Vermont, Dept. of Microbiology and Molecular Genetics, Burlington, Vermont

<sup>2</sup> Université de Montréal, Dept. d'informatique et Recherche Opérationnelle, Montréal, Québec, Canada

The hairpin ribozyme, assembled with a short substrate RNA, presents two loop domains (termed A and B) bounded by short helical regions. Most of the catalytically essential groups are concentrated within loops A and B and the two independent folding domains of the hairpin ribozyme must interact for the ribozyme to be active. Computer generated three-dimensional models (MC-SYM) using constraints from photoaffinity crosslinks and hydroxy-radical protection data, have shed some light on the relative spatial arrangement of the A and B domains, including the existence of an essential interdomain base-pair (Pinard et al., 1999). High resolution structures of the isolated domain based on NMR spectroscopy are also available. However, little is known about the interdomain contacts or the internal geometry of the individual domains, which are usually depicted as straight structural elements. We have investigated the internal geometry of the domain A (the substrate binding domain) in a catalytically active complex. Domain A comprises two short helices (6 and 4 bp) that flank a small symmetrical loop (4 nt in each strand) containing the cleavage site. Using circularized substrate and various constructs of the hairpin ribozyme including a circular substrate binding strand we have shown that domain A adopts a bent conformation. Our data indicate that circular substrates are cleaved very efficiently and that a circular substrate binding domain is catalytically competent. These data strongly suggest that the bent conformation reflects the actual structure of an active complex. These observations differ from the conformation obtained by NMR, and provide new three-dimensional constraints for the understanding of the hairpin ribozyme structure and function. Using these topographical constraints in combination with our previous biochemical data and computer assisted molecular modeling (MC-SYM) we have generated a model describing the new spatial arrangement of the substrate binding domain.

## Basic geometric problems on reconfiguring polymers

Michael Soss and Godfried T. Toussaint  
School of Computer Science, McGill University

We examine a few computational geometric problems which may aid the determination of whether or not a polymer can be reconfigured from one folding to another. We use an off-lattice model of a polymer, a polygonal chain (path of line segments) in three dimensions. The polymer can be reconfigured in any manner as long as the edge lengths and the angles between consecutive edges (bond lengths and bond angles) remain fixed, and no two edges intersect during the motion. We discuss preliminary results on the following problems.

Given a polymer, we select some interior edge, defining the two subchains adjacent to it. We keep the two individually rigid and perform a dihedral rotation on one subchain at the specified edge while leaving the other fixed in space. The motion is deemed to be feasible if the polymer does not self-intersect during the process. As the polymer is locally changing only at a single vertex, this is the simplest motion allowable in our model. We call this motion an edge spin. We give algorithms for determining the feasibility of these motions and prove that they are near-optimal.

In determining whether a polymer can be reconfigured from one folding to another, it is useful to consider reconfiguring through some canonical conformation. In our three-dimensional case, the most obvious choice is to flatten a polymer into the plane. However, we demonstrate that determining if a given polymer can be reconfigured into the plane without self-intersecting is NP-hard, even if the restriction that it must lie monotonically is added. We provide additional tools such a linear-time algorithm to decide if a polymer has a non-crossing convex coil conformation (where all angles turn in the same direction), although deciding if a sequence of motions to reconfigure it into a convex coil remains an open problem.

## Linguistic analysis of the 5S ribosomal RNA of Actinomycetes

Sandra M.R. Subacius and Gabriel Padilla

Departamento de Microbiologia, Instituto de Biociências, Universidade de São Paulo (USP)

CEP:05508-900, São Paulo, SP. Brasil

The Actinomycetes group consists of several genus of saprophytes and simbiotic-parasite gram positive eubacteria, that inhabit the soil and also the natural cavities of man and animals. These bacteria have a great morphological diversity, extending from the coccus (*Micrococcus*) to genera with a permanent and highly differentiated mycelium (*Streptomyces*).

Many of the Actinomycetes species exhibit a mesophile nature, nevertheless their genomes are rich in G+C. The 5S-genes, like other regions of the genome, show G+C composition also shifted from the neutrality (50%). As consequence, the 5S rRNAs of the Actinomycetes present G frequencies (or C in *Mycobacteria*), higher than those observed for the thermophilic eubacteria. However, the G and C frequencies in the Actinomycetes have been kept high by the preferential fixation of point mutation of the transversion type. This behavior, as expressed by the Ks index (see Subacius and Bussab, 1998) involves unbalanced exchange between the G (purine) and U (pyrimidine), and it has isolated the Actinomycetes from other eubacteria.

In order to know the implications of the high G+C contents in the variability of the di and trinucleotide vocabulary, a linguistic analysis (Brendel et al. 1986, Trifonov 1988, Pietrokovski 1989) of the Actinomycetes 5S sequences was carried out from data available in the EMBL Data Bank. According to this method, each oligonucleotide word in the vocabulary receives a markovian contrast value, that can be either positive (preferred word) or negative (avoided word). Oligomers consistently over or under-represented among different sequences, would be clearly good candidates for functionally important sites.

The results of this research have shown that, although the Actinomycetes are distant from other eubacteria either by the high G or GG contents, the CC dinucleotide was the preferred word in 90% of the species, with positive contrast values higher than stds  $\gg 2$  (standard deviation), while its complementary counterpart, the base-pair GG, remained in the expected interval. The CC dinucleotides are mainly concentrated in the hairpin C of the 5S secondary structure. This region is also characterized by the occurrence of CCG trinucleotides (preferred word), that overlaps to the CCC triplets, making up the CCCG tetramer, located in the strands 5' and 3' of the helix III. The CCCG tetramer seems to be in fact the word with local functional meaning.

The std-values obtained for the CCC trinucleotide were consistently negatives indi-

cating that CCC, contrary to CC dinucleotide, tend to be a word avoided in the contrast vocabulary of the Actinomycetes 5S molecules. In 92% of the Actinomycetes GGC was strongly avoided. The occurrence of the GGC sites was scarce along the molecule circumscribed to 1 or 2 sites (in tandem) in general in the 5' strand of the helix I. That region, the presence of two sites, strategically occupied by one of the two pyrimidines (U and more rarely C) prevents the formation of an G cluster. The behavior observed for the trinucleotide GGC, reflects an intense restriction on the GC dinucleotide, except for the 5S of Mycobacteria. Another important difference that set apart Mycobacteria from the Actinomycetes, is the site GAA (preferred word in 100% of the Mycobacteria species), overlapped to an AAAA site (GGAAAA) in the loop E. The A cluster found only in the 5S of the Mycobacterias among the Actinomycetes constitutes the signature of the group. Despite the GC dinucleotide being an avoided word, the AGC trinucleotide, the third most preferred word in Actinomycetes, is found in 50% of the species. AGC sites appear in four positions in the molecule, in the strands 5' and 3' of the helix II (isolated sites), in the loop B (overlapped to the AACG tetramer), in the 3' strand of the helix III. This latter position is very interesting due to the A bulges in the position 50, that represents the central base of two preferred words in the Mycobacteria vocabulary (GAA and AGC) that overlap to the GAA50GC pentamer. That region is conserved in near all the Actinomycetes species.

Brendel, V. Beckmann, J.S. and Trifonov, E.N. (1986) Linguistic of nucleotide sequences: morphology and comparison of vocabularies. *J. Biomolec. Str. Dyn.*, 4:11-21.

Pietrokovski S. (1989) Nucleotide sequence dialects and vocabularies. MSc. Thesis, The Weizmann Inst. of Science, Rehovot, Israel.

Subacius, S.M.R and Bussab, W.O (1998) Purine and pyrimidine composition in 5S rRNA and its mutational significance. *Genetics and Molecular Biology*, 21:255-258.

Trifonov, E.N. (1988) Nucleotide sequences as language: morphological class of words. In Block H.H. (eds) *Classification and related methods of data analysis*, Elsevier-North Holland, pp.57-64.

Acknowledgements- FAPESP-Brazil

## Mining the databases for mobile group II introns

Steven Zimmerly, Georg Hausner, and Xu-chu Wu  
Department of Biological Sciences, University of Calgary

Mobile group II introns were identified in the databases based on the reverse transcriptase (RT) ORF typically encoded within the introns. Over 60 group II introns were compiled including 40 mitochondrial, 11 chloroplast and 18 bacterial ORFs. Nearly half of the bacterial ORFs were not previously identified as group II introns. The bacterial group II introns are invariably found in mobile DNAs or plasmids, are sometimes located outside of functional genes, and are frequently found as truncated fragments. Together the data suggest a higher degree of intron movement and flux in bacteria than in organelles.

A phylogenetic tree of the compiled ORFs was constructed to predict a history of mobile group II introns. The phylogenetic tree was rooted with RTs of telomerase, non-LTR retroelements and retrons, and the inferred phylogeny shows two major clusters of group II intron ORFs. The mitochondrial cluster contains nearly all known mitochondrial group II intron ORFs, while the chloroplast-like cluster is a heterogeneous collection of all known chloroplast group II intron ORFs plus algal mitochondrial ORFs and bacterial ORFs. A number of bacterial ORFs do not have affinities to either major lineage and are probably early branching. The data give an overview of the exceptionally high degree of horizontal transfer of group II introns, mostly among related organisms but occasionally between organelles and bacteria. The Zn (nuclease) domain was lost multiple times during evolution, and is lacking from many bacterial ORFs. Overall the data are consistent with a bacterial origin for mobile group II introns.

