Research proposal

December 13, 2004

My research proposal focuses on computational study of RNA structure and functionality. More specifically I will use the following subjects to be the skeleton of my work plan: 1) Finding proper representation for the structure. 2) Developing data mining tools to connect structure and function. 3) Developing algorithms to predict local structures. 4) Writing efficient codes to follow dynamics within small subunits of RNA. 5) Analyzing and predicting structure-mediated interactions of RNA and proteins. Lastly I present a possible application to Retroviruses encapsidation processes.

The research plan to cope with these challenges is detailed below.

RNA structure and functionality

RNA is a multifunctional biopolymer that takes part in transforming genetic information from DNA into a well defined set of proteins. Genetic messages are imprinted in RNA nucleic acids sequence via transcription from the DNA. Protein synthesis, performed with the aid of the tRNAs on the ribosomes, is involved with catalytic and recognition processes. In other processes such as self splicing RNA functions as an enzyme. On top of this enzymatic activity, RNA has roles in structural activities (templates for virus capsidation) and even appears to participate in protein transport.

This variability in functionality depends on the ability of RNA to fold into a large number of exact spatial forms. The folding process is dominated at the first stage by base pair interactions that form a secondary structure which is sequence dependent. In a second stage, hydrogen bonds as well as stacking interaction arrange separate regions of the secondary structure to form large

scale three-dimensional structures. At the microscale the conformations are dictated by complicated interatomic potentials. An exact calculation of the potentials is impossible by present days tools and some simplifications are needed. The conformation of a single residue in RNA depends mostly on a set of torsional angles: six for the backbones, one for the base orientation and a sugar pucker. RNA structure is based on a flexible backbone with a small set of four bases that are not greatly different in structure. Nature uses a different design concept for proteins, where each residue has only two back bone torsion angles, but has great variability in side chain chemistry. We have shown, with a simple model for torsional potential, together with statistical information from x-ray structures, that the conformation of a single residue can be reduced to a limited number of states. We refer to these states as "bins" and attach an ascii code letter for each bin. With this we are able to represent RNA structure as a text. For example, we chooses the letter "a" to represent a conformation of a residue in an "RNA A Helix form". Then a strand with all the residues in this conformation will be represented by the word "aa...a". In another example, a tetraloop is defined by four residues capping a double helix. The second residue has a turn in the back bone that form a conformation we annotated as "o". The GNRA tetraloop motif is represented then by the word"aoaa". The text representation simplifies the complex problem of recognizing 3D motifs into text editing and processing. This method was found to be exact when used to recognize short familiar structural motifs.

My plan is to improve and then to fuse this coding method into other single residue qualities, such as type, charge distribution and connectivity, to form a large data mining tool that can be used for analyzing and predicting RNA structure and functionality. More specifically, The plan of research includes:

1. **Single nucleotide conformation:** Single RNA nucleotide conformation is defined by the aforementioned set of seven torsional and one sugar pucker, that can be reduced to six backbone torsional angles. We have discretized the conformation to bins based on statistics over crystallographic structures in the way described above. We have used tools from cluster analysis and image processing such as k-means to improve and automatize the classification. Thus far those results are less accurate than our manual binning method. I will use other tech-

niques to find a complete set of clusters, each one representing a group of conformations that are easily approached (low energy barriers) from other conformations in the clusters. One possible way is to use a Hidden Markov Model. Indeed there is a correlation between back bone torsional angles of two adjacent residues at their edges due to sterical reasons. This correlation decays quikly with the distance hence a Markov model where the residue number is the "time" parameter is reasonable. The states of the models can be any spatial partition of the torsional space. The observation will be a multiple Gaussian distribution within the states. The reassesment formalism will be used on different initial models to find a transition matrix and best distribution for clusters. Another mode of action is to analyze charge distribution for different conformations and to cluster them according to this criterion rather than the spatial one.

2. **Fusion of three dimensional coding with secondary structure:** The previous description takes the RNA to be a unidirectional sequence, that fails to describe base pairs interactions. Back bone torsions do not define the geometry of base pairs. With this backbone description we lose some important information encoded in the secondary structure. I will develop a visual representation that will include the structural alphabet in a secondary structure representation. A binning model that will include secondary structure will be developed with the aid of extra partitioning to paired and non paired residues. With this new language as well as with the use of graphs, I will look for repeating structural motifs, that are defined by backbones and base pairs, in the RNA. A combination of similarity in the local conformation and of the array of hydrogen bonding will give a suitable definition for a structural motif because such a structure is defined by a high thermodynamic stability.

3. **Interactions with proteins and ions:** From crystallographic data it can be seen that a large portion of the ribosome residues are exposed to interactions with water, with ions and with proteins. This is true for every functional RNA in the cell. The interactions are mostly via the bases and the phosphate group in the backbones. From the data analysis it seems that first shell interactions with phosphate oxygens

play a major rule in immobilizing cationic groups by the RNA. Such interactions are highly correlated with specific geometries. This correlation hints about an existence of ion or amino acid binding motifs. I plan to derive a method to find these motifs with high degree of reliability. The problem is not straightforward. A single amino acid, for example, can interact with more than one nucleotide. These interactions are not necessarily confined to vicinity in sequence or even within the secondary structure. Some of the interactions can be obscured in the crystallographic structure. For example, interactions via hydration shell are not always visible in a crystal structure. New data mining tools must be developed to single out such possible motifs. One way to deal with this task will be to group sets of RNA residues affiliated in some manner with the same amino acid (or peptides). Such sets are known from other data mining fields as "transactions". At the next stage, "frequent items sets", i.e. sets of residues that support large enough number of transactions should be found. Then association rules have to be determined on these sets. Lastly these rules will be used to devise a partition of the residue sets into different amino acid (peptides, ions) binding motifs.

4. **Prediction of constrained small portion of RNA:** Crystallographic structures of most RNA molecules are not available. Most structure determinations of RNA do not have high enough resolution to define its conformation by torsional angles. It is comparatively easier to get the secondary structure with prediction methods that are based on the sequence. It is also more feasible to get accurate crystallographic structures for small substructures of a large RNA. The possible conformations of a single residue in RNA are limited by external constraints such as hydrogen bonding or spatial confinement. Given a set of these constrains it is possible to use inverse kinematics to calculate all the possible conformations. This method can be extended to a sequence of residues. I intend to use this method to analyze and to predict the 3D structure of short loops that seem to be highly confined by the stem and by hydrogen bonds. At the next stage, this procedure will be used for short portions of the RNA which are associated with proteins. Dynamical aspects of transitions between structures will be represented by thermodynamical rates. A special stochastic integrator

will be developed to give an efficient dynamical tool.

5. **Retroviruses encapsidation:** A possible application for the described methods is analysis of retrovirus encapsidation processes. Retroviruses are RNA viruses. Their genetic information is stored in a relatively small RNA. The compactness of the viral RNA allows it to carry information for only a limited number of proteins. Some of those are used to produce a shell or a capsid that protects the fragile RNA in the cell and in the intercellular environment. The economy in information forces these capsid proteins to be homogeneous. This dictates a high degree of symmetry for the capsid shape. The natural closed shapes are tubes and icosahedral. There are variations within these two classes but every virus will have its specific shape. This is not a one to one description because some viruses such as lentiviruses can be polymorphic and the capsid will change its shape during maturation.

   There is evidence that the encapsidation of a virus is a self assembly process of the capsid proteins units on a RNA template. Indeed it is important that the capsid subunits will package the specific viral RNA and not the cellular RNA, or just assemble to empty shells. The recognition process is far from being understood for several reasons. First, there are only few crystal structures of RNA inside a capsid. Secondly, even in those, most of the structure is obscured. Because of this, only partially secondary structure results are available. Secondary structures of the RNA show which regions that alter the encapsidation efficiency have a complicated stem loop structure and a large portion of the interactions with the capsid protein are via single strand regions, where the geometry is variable. Strong protein-RNA interactions are salt bonds with the phosphate groups. these kind of interactions show high correlation with unusual backbone geometry. Indeed, encapsidation of the Tobacco Mosaic virus forces the single strand RNA into an extended geometry, such that salt bonds are formed between the RNA phosphate and amino acids, mostly Arginine.

   Understanding the encapsidation process has an enormous practical implementation. Drugs can be developed to tamper with the structure of the viral RNA in regions which are crucial to the virus assembly or disassembly. There are no complete and high resolution crystallo-

graphic structures of latent retroviruses such as HIV. Achieving such structures is still a challenge. Even after solving a unique structure, the encapsidation is a dynamic process, which may involve more than one intermediate structure. Drug mediated structure changes are another process that demand more 3D data that cannot be found without laborious or even impossible crystallographic measurements.

Methods such as those I propose to develop seem to be reasonable to deal with these problems. I will look for the 3D structure of the RNA in regions that interact with the capsid proteins. An effort will be made to reconstruct these regions. I will use possible peptide binding motifs as corner-stones for possible substructures. Then I will use design methods such as inverse kinematics to deduce larger scale regions in the viral RNA. The proposed possible structures will be tested by comparing it to similar sequences and secondary structures in other RNAs with known 3D structures. At a more advanced stage, I will use existing crystallographic measurements as well as structures that I will achieve by collaboration with research work of crystallographic lab, to validate and complete the model. At the final stage, models that depend on RNA structure and on RNA-amino acid interactions will be used to develop descriptions of the dynamics of the templated encapsidation. In another related project I will look for structure changing interactions within RNA such as intercalation to give a basis for therapeutic methods.

1 Past research accomplishments

My work in the past has been on dynamics of microscopic systems coupled to macroscopic environments. Within the framework of this subject, I worked on the following projects:

1. **Multidimensional reaction rate theory**: Reactions of chemical systems that are activated by a heat bath have been analyzed by projecting the dynamics on a stochastic equation (Langevin equation) describing the evolution of a reactive degree of freedom in the presence of random noise and friction. I have extended this model to give reaction rates of multidimensional systems coupled to a heat bath. Through this project I developed skills in solving stochastic differential equations and writing codes to integrate large molecular systems.

2. **Surface diffusion**: The dynamics of atoms on crystal surfaces is formulated by deterministic potential that is calculated from the equilibrium configuration of the lattice and random terms describing interaction of the adatom with the crystal phonons and electrons. I have used this model to analyze the diffusion of metal atoms on metal surfaces. The model we developed explains the anomalous dependence of the diffusion measured for Pd atoms on W(211) surfaces.

3. **Thermal surface diffusion controlled by external field**: Thermally activated system driven by a weak external AC field can show great enhancement of reaction rates. We have found an analytical solution for such a problem in the weak friction regime.

4. **Molecular motors**: An external field periodic in time acting on a thermally activated system is known to produce directed currents in the systems. We worked on a problem of such rectified currents on a lattice in the weak friction regime. We analyzed the problem numerically and suggested an analytical solution.

5. **Electron transfer process in a conducting polymer**: We worked on a classical description of electron transfer from the donor to the acceptor over a molecular bridge. We have utilized the theory of activated rate processes to find an expression for the electron transition

rate. The long range purpose of this work is to develop a model for electron transfer in organic polymers such as DNA.

6. **Fast numerical integrator for stochastic equation of motion**: Numerical solutions of stochastic differential equations such as the Langevin equation are generally time consuming. The existing codes have a low degree of accuracy and are generally unstable. I have developed a new class of algorithms with a high degree of accuracy. With these codes I was able to find numerical solutions for problems that were not solved before.

7. **Dynamics within system coupled to anisotropic heat bath**: In this project we developed numerical tools to integrate the time evolution of an object in a time and space dependent bath. We assume a general model where memory effects in the bath were taken in consideration.

8. **Reaction and diffusion in nematic liquids**: I have developed a simplified model for dynamics of a particle in an environment of lyotropic liquid. I have shown that by using a Magnetic AC field thermal processes can be greatly modified.

9. **Conformational analysis of RNA structures**: I have developed a classification method for structural motifs in RNA called "binnin". I have written a code that enable fast recognition of these motifs.

Teaching Statement

My teaching experience includes:

- **Classroom Teaching of the course:** *"Linear Systems and Control" in the School of Electrical Engineering of Georgia Tech.* This is a graduate course that is mathematically oriented. I prepared and taught the mathematical part, which was about two third of the course. This course gave me experience in teaching in front of a relatively large class (about 80 students). Part of my duties were to teach the course to distance learners. This program gave me experience in teaching using telecommunication.

- **Facilitating in the course:** *"Problem Based Learning" in the School of Biomedical Engineering in Georgia Tech.* In this course groups of students presented solutions to given biotechnological problems. As a facilitator, I introduced the subjects and supervised the students in solving scientific problems.

- **Guiding graduate students:** I am currently guiding two PhD students. I am teaching them some specific elements from Biochemistry and Biophysics needed for their work as well as Bioinformatic programming.

I find teaching to be one of the most rewarding parts of scientific work. On the scientific level, explanations and transfer of ideas improve the teacher understanding of any discussed subject. Moreover, on the personal level, shaping students ways of thinking and stimulating originality, are by themselves great award.

My teaching philosophy is based on the assumption that the student has the potential to learn. Then the teaching should maximize the ability of all the students to develop this potential. The goals of teaching are to provide the student with concepts and tools to treat scientific problems independently and in a creative manner.

To achieve these aims the following means can be used: First the teacher needs to understand the material to a high degree. Then he has to edit the classes into cohesive studying units. For each unit objectives should be stated and then methods and concepts should be built around it. The teacher has to monitor the progress of the class and have a number of alternative explanations in case when the material is not understood. Examples are an important part of the teaching. They should be brought both from the field of study and from other related fields. I found examples from modern study case or from more advance material to be a focus of interest and inspiration for students and for the teacher. The teacher should be available for extra explanation after class hours. This is an important part of the course that helps to correct misunderstanding on a one-on-one basis . It is also valuable for the teacher as a direct measure for understanding of the class. The same is true of the homework assignments. In there, challenging conceptual questions should be used among practicing exercises. More advanced courses, will include projects and guided reading part to prepare the student for an independent scientific work.

My background in Physics, Chemistry and Mathematics seems to be suitable to teach courses such as: Fundamental Physics, Chemical Physics, Biophysics, Numerical Analysis, Mathematical Methods for Biology, Computational Biology, Probability, Stochastic Methods and Bioinformatics courses.