A mechanical overview of RNA in an energy focused world - Draft rev1
Author: Jennifer Pearl
Date:5/20/2022

Chapter 1: What is the ensemble?

I have wanted to write this for a very long time, but the thought of writing a document or a book on how I understand RNA to behave at this point in time after now being published as a contributor and to have my work cited in reference to my research in RNA design methods in a peer review journal and seeing my many theories proven right through *in silica* and *in vivo* testing. To start with, I truly believe that RNA is a mechanical system that lives in an energy world, and thus us bound by the rules of mechanics of materials more than the effects of the energies in the loops. I believe that this is why riboswitches are so difficult. When a switch is made there are shifts in energies and the hydrogen atoms that bind the nucleotides together are pulled apart and pushed together somewhere elses they have a strong attraction. That shape they like to take is what is called the MFE or minimum free energy state and it is the one that a RNA design likes to hold most of the time. Now read that most of the time as “usually kinda but not really when you think of it”. This is because the MFE structure is only one data point in what is called the ensemble.

The ensemble consists of all the possible pairing configurations of the nucleotides in the primary structure and thus it consists of all the secondary structures that the single primary structure can take. As you move further from the MFE or lowest theoretical energy state you can get, the number of possible secondary structures increases. This is because the overall energy state of the design has changed so teh originally predicted pairing probabilities of the base pairs has shifted along with the calculations and now pairs that were strong are just too weak or those that were originally were too weak are just now strong enough to pair up, and thus predicted secondary structure can be different by orders of magnitude or just slightly. Now with this in mind, the overall energy state of the RNA is calculated with using temp as a variable at time of calculation. It is impossible to keep a perfect temperature and thus any fluctuations will cause energy shifts. There are many things that can tug and pull on that MFE state, in a constant fight to pull the MFE away from its cozy low energy home out into the busy world of higher energies.

Now…with all that in mind and taken into consideration I hope my response of “usually kinda but not really when you think of it” makes sense. When you think of this if you focus all your energies on the MFE or just energies you may miss something. It is more of a kaleidoscope of what could happen, than a dial in type thing. You have to think of it in terms of at any one time all the RNA nucleotides are tugging and pulling on every other molecule. It does not matter how far away they are as there are always tiny forces. These tiny forces are represented in teh partition function as the statistical probability that two nucleotides will bond and that is what is referred to as pairing probabilities.

Pairing probability is a term used to refer to statistical calculations performed that represent the statistical probability that specific nucleotides will pair up and form a base pair. This does not

care if a base pair is a Watson-crimes canonical base par or whatever. I tend to think of it as the amount of attraction and pull each nucleotide has on each other in the entire primary structure.

These pairing probabilities then guide and really determine the secondary structure of the RNA based on the energies of the bonds whose forces are represented by the pairing probability's. I'm not really sure at what point energys come into play In my mind, but in general energies are defined by the type of base pair involved as well as the base pairs and unbound nucleotides nearest it. It is with certain geometry's that you then get energy bonuses such as special loops, stacks, binding sites, etc, this includes boosting. In this thought process, boosting is adding a nucleotide in a location that causes it to have a pull on other RNA and this increasing it's strength of pairing probabilities and attractions at certain locations to help achieve a desired shape. This might be what is actually creating energy bonuses. It's the rigidity of the structure and the high pairing probability is a result of string energies. It is a bit circular, but only to a point.

The application of analysis of pairing probabilities is an interesting thing that has been the focus of the majority of my research for the entirety of the project. It was the analysis of the pairing probabilities calculated by NUPACK and Vienna 2 that resulted in all the winning designs submitted by Sara during the DOE. There are a few ways to look at the data.

Chapter 2: Analysis of Partition Function: Paring probabilities and the ensemble

My research has shown that the partition function is probably the most important aspect of an RNA design yet it is not a very well understood thing. The most important aspect I think is the ensemble, as the only standard metric that is currently part of Vienna2 and NUPACK and this directly citing a paper that I have ever seen track with a design score is ensemble diversity or ensemble defect. In fact when I first noticed this back in 2016, I did a search into the old Etrna forum posts and found people pointing this out around 2011, but the post I found was largely ignored… I have never seen any other metric, such as MFE, centroid energy, whatever actually contains a signal that something is happening that is not a completely unique problem that changes from design goal to design goal. This is because all the RNA sequences are unique and there will never be ideal energy levels I believe, but there are ideal mechanical configurations and ensemble diversity's signal I believe is based on its relation to how stable the secondary structure is and how much diversity there is the ensemble.

Now why start a chapter on pairing probabilities with ensemble defect rant? The answer is that pairing probabilities essentially determine the ED and the ED really points to a RNA's ability to fold right and/or just be stable. My research shows time and time again a clear slope in all the plots across all the labs that ED has a signal that clearly shifts as the Eterna score and fold change go up, regardless of the lab. The thing though is that each lab seems to have a bit of a different ED range that is good for each lab but they are all kinda close.

What are these pairing probabilities that are so important and how do they work? Now remember that RNA is bound by nucleotide pairs whose bonds are generated by hydrogen atoms and the number of atoms is the strength of the bond. The number of hydrogen atoms is determined by the nucleotide type and each type has a specific number with each Watson-Cricks base pair having one of four hydrogen configurations. The weakest pair, AU

When i think of pairing probabilities i think of the dot plots you can look at in Eterna as those are a graphical representation for the pairing probabilities data. A dark spot is a strong predicted bond and a light spot is a weak pairing and a white spot is no pairing. The data outputted from NUPACK and VIENA2 is formatted as numerical values representing the statistical calculation of chance to bond. It is normalized to the value of 1 with 1 being a very strong bond abd the smaller the number gets the weaker the bond abd likely to pair. The models output a list of every potential pair and probabilities for those pairs. That is what you are seeing when you look at the dot plots as well as the cool new tool arcplot, in that they are a graphical representation of those attractions and bonds.

To get back on point and how this all applies to the ensemble, the ensemble can be thought of as all the possible secondary structures that a design can take if you consider the entire range of possible kcal's and not just the MFE which is just the edge of the ensemble. Since the probabilities are just a statistical probability there are chances that the design will have slightly different results on occasion, not large mind you, but enough to nudge the design. When we get

the MFE secondary structure we are then looking at the most probable shape it will take at the lowest and thus strongest energy levels that can hold the RNA strong. You can however, peek at what things would be like if the energy levels were not perfect and that is the subopt function of NUPACK. With this you can get a list of secondary structures in the ensemble found at a specific kcal delta as well as find out how many possible secondary structures there are. Remember I said a few times that pairing probabilities define the secondary structure, well you can see that in action now and understand what is happening.

Let's walk through this process one last time now that we know what is going on in the steps.

A RNA sequence is the most stable at its Minimum free energy state (MFE) and naturally wants to stay there when at the temperature the MFE was calculated for. This is because as RNA gets colder the bonds become stronger and as it gets warmer the bonds begin to fall apart (this is what the melt point metric shows). Now it's impossible to keep a molecule at a specific temperature perfectly and there are always numerous environmental factors affecting RNA, so the actual energy levels fluctuat a little. When the RNA fluctuates it's energy levels the secondary structure vary depending on the delta from MFE, abd the design will fluctuate between all the possibilities secondary structures in that kcal delta group of the ensemble, abd if it goes back to the MFE it will be back at the MFE secondary structure. Now some designs lend to very few different alternate secondary structures and some lend to a lot of alternate secondary structures in each unique group. Some of the groups will have very little change or shift from the MFE secondary structure and some will have a lot of change or shift between the secondary structures, regardless of how many secondary structures there are. The less variation and the less alternate secondary structures in each group of kcal delta's the more stable the design will be then. This leads into another question…why is it more stable when there is less variation? That answer is given I believe by mechanics of materials for both dynamic and static loads.

Chapter 3: Why is mechanics of material so important if it’s not till chapter 3? Also, what is mechanics of materials?