1-1-2011

Artificial and natural nucleic acid self assembling systems

Marcus Wood
Wayne State University,

Follow this and additional works at: http://digitalcommons.wayne.edu/oa_dissertations

Part of the Biophysics Commons, Chemistry Commons, and the Nanoscience and Nanotechnology Commons

Recommended Citation

ARTIFICIAL AND NATURAL NUCLEIC ACID SELF-ASSEMBLING SYSTEMS

by

MARCUS WOOD

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2011

MAJOR: CHEMISTRY

_____

Advisor   Date

_____

Advisor

_____

_____

_____

DEDICATION

To my family, friends, and teachers, throughout the years, without whom I never would
have finished my degree.

ACKNOWLEDGEMENTS

There are too many people that have helped me through the years to acknowledge individually. I am grateful to them all, and so my degree is the product of an intellectual community, and not solely one person. However, I would like to pick a few to single out. I would like to thank my committee, my co-advisors Drs. SantaLucia, and Rueda, and my other committee members, Dr. Cunningham, Dr. Poole, and Dr. Chernyak. I would like to thank Dr. SantaLucia for many stimulating scientific discussions, a good grounding in nucleic acid biophysics, guidance throughout my Ph.D. thesis, and editing this entire document. I would like to thank Dr. Rueda introducing me to the single molecule world, giving me an appreciation for publication quality presentations, and for providing the TRAP project and space in his laboratory so that I could have a publishable project and complete my degree. I would like to than Dr. Cunningham for keeping the biological side of my projects in balance with the physical, which had a tendency to become ascendant. I would like to thank Dr. Poole trying to impart to me the skills and attitudes a competent analytical chemist should possess; a deep respect for statistics, data reduction, and the importance of robust, precise, and accurate measurements and methods, topics that are often neglected by the other divisions in chemistry. I am very grateful to Dr. Chernyak for being willing to join my thesis committee at an extremely late date and with very little notice, and helpful career advice.

I would also like to acknowledge our various collaborators. Many thanks are due to Dr. Sergey Malinin for the development of the mathematics for the statistical model of TRAP binding, and many interesting and helpful discussions. Thanks to our collaborators Drs. Foster and Sachleben, for supplying the TRAP protein and encouragement. I am

defender, and her great patience with the unreasonable amount of time my degree has taken. At last, I would like to say thanks to the entire complement of the Rueda and SantaLucia labs past and present; you have all been very good friends, and have helped to keep me from adding felonies to my resume.

Finally, I have found some comfort the in excellent poetry from World War One. The authors speak very movingly about hopelessness, abysmal leadership, broken ideals, heartbreaking naiveté, nationalism, misplaced trust, loyalty, honor, and duty, and the death of an entire generation of young men. An interesting coincidence between my work and the conflict is that I worked with mechlorethamine, a chemical derived from chemical weapons research on the mustard gas used in World War One. The effects of this chemical on lymphoid tissue suggested its use on cancer and success with this effort started the entire field of cancer chemotherapy. It is a horrifying irony that millions of lives have been saved by compounds similar to those that have caused so much suffering. Below are two examples that I have found particularly relevant. The authors of the poems, Siegfried Sassoon and Wilfred Owen, are some of the better poets in the English language and should be remembered during our current wars.

Suicide in the Trenches

I knew a simple soldier boy.....
Who grinned at life in empty joy,
Slept soundly through the lonesome dark,
And whistled early with the lark.

In winter trenches, cowed and glum,
With crumps and lice and lack of rum,
He put a bullet through his brain.
And no one spoke of him again.

You smug-faced crowds with kindling eye

Who cheer when soldier lads march by,
Sneak home and pray you'll never know
The hell where youth and laughter go.

Siegfried Sassoon - 1918

DULCE ET DECORUM EST

Bent double, like old beggars under sacks,
Knock-kneed, coughing like hags, we cursed through sludge,
Till on the haunting flares we turned our backs
And towards our distant rest began to trudge.
Men marched asleep. Many had lost their boots
But limped on, blood-shod. All went lame; all blind;
Drunk with fatigue; deaf even to the hoots
Of gas-shells dropping softly behind.

Gas! GAS! Quick, boys!—An ecstasy of fumbling
Fitting the clumsy helmets just in time,
But someone still was yelling out and stumbling
And flound'ring like a man in fire or lime.—
Dim, through the misty panes and thick green light,
As under a green sea, I saw him drowning.

In all my dreams before my helpless sight
He plunges at me, guttering, choking, drowning.

If in some smothering dreams you too could pace
Behind the wagon that we flung him in,
And watch the white eyes writhing in his face,
His hanging face, like a devil's sick of sin,
If you could hear, at every jolt, the blood
Come gargling from the froth-corrupted lungs,
Bitter as the cud
Of vile, incurable sores on innocent tongues,—
My friend, you would not tell with such high zest
To children ardent for some desperate glory,
The old Lie: *Dulce et decorum est*
Pro patria mori.

Wilfred Owen - 1920

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1. Introduction

Nucleic acids are good candidates for nanomachine construction. RNA, along with its close chemical cousin, DNA, participate in most of the processes necessary for life (information storage, catalysis, regulation, protein synthesis, etc.). Many of these processes are in effect, nanomachines, selected by evolution to perform a task important to the cell. The ribosome is a wonderful example. It is an exquisitely efficient and reliable molecular factory that takes information (from the DNA in the form of an RNA transcript) and produces the desired protein. This process is accomplished with an error rate that would be the envy of a modern intensive care unit (0.0006 errors per amino acid incorporated in a polypeptide chain vs. 0.01 errors per activity performed on a patient in the ICU![1]). The catalytic center of the ribosome is RNA; the ribosome is a ribozyme (i.e. a nucleic acid playing a catalytic role). There are many other examples of nucleic acids acting as nanomachines by themselves or as part of larger architecture (e.g. group II introns[2], chromatin[3], siRNA[4], etc.). With this inspiration from the natural world acting as goal and blueprint, it makes sense to use nucleic acids as the building blocks of nanomachines. This inspiration is not the sole reason for using nucleic acids as a framework for nanotechnology. The work of the past many years has given us a good understanding of nucleic acid synthesis and function. Short nucleic acids can be chemically synthesized by robots, and longer ones can be made using polymerases from DNA templates. Even longer strands can be pieced together with enzymes called ligases or recombination[5]. The thermodynamics of the association can be predicted, allowing us tune the association of the parts of a nanomachine[6]. Cellular replicative machinery can be adapted to replicate nucleic acids, allowing mass production of a particular nanomachine.

DNA is a relatively stable molecule, as evidenced by its use for genetic material, suggesting its use as a longer term construction material. RNA is relatively more unstable, suggesting its use as scaffolding material, or for degradable or recyclable nanomachines. Many enzymes that deal with processing RNA and DNA are known. They can make cuts in designated places, or bring pieces together. They can perhaps act as construction machinery for a designed nucleic acid nanomachine, saving the trouble of constructing the tools needed to build a nanomachine.

What is needed to actually bring complicated machines to reality? The necessary components are design, deep understanding of material properties, and dynamics. A complicated structure, or machine is no longer designed by hand, it is designed in a computer with purpose built software. This is due to the complexity of the product, and the process of creating a workable creation. The same holds for deigning using nucleic acids. While some useable design software exist[7], there is a need in the community for a reliable software tool to help in the design process. An architect does not design a building without a deep understanding of the properties of the structural materials used and the design principles involved. A good understanding of nucleic acids has been wrested from nature by the work of countless molecular biologists, but mostly in the context of living systems. Some of this information is directly applicable to nanoconstruction, but more basic work must be done to understand nucleic acids as construction materials, so that the nanodesigner can have the same confidence that an architect has working with steel, concrete and the arch. Lastly, we are truly interested in nanomachines, not nanostructures or nanoarchitecture. This implies motion or action. The start of design may be static, but the end is an active machine that can accomplish work.

For this a better understanding of the dynamic nature of nucleic acid must be acquired. This work will try to address in a small way each of these questions, a better computational tool for design, a deeper fundamental understanding of the mechanical, structural and geometric properties of nucleic acids by attempting to construct a nanostructure, and finally, to start probing the dynamic properties of nucleic acids as inspiration for a dynamic machine.

Chapter 2. Nucleic Acid Force Field Development

2.1 Introduction

To start, the details of nucleic acid structure still need to be better determined. There are far fewer crystal structures of nucleic acids than of proteins. The Protein Database (PDB) contains 60769 released structures for proteins, but only 4858 released structures for nucleic acids alone and as complexes with proteins as of 6/14/10. Computational methods for predicting nucleic acid tertiary structure lag behind protein methods[8], partly because RNA was experimentally harder to prepare than proteins, and partly because the biological role that RNA plays was not fully appreciated until recently. Complicating the matter, RNA has six backbone degrees of freedom per monomer, compared to only two for protein, making the challenge of structure prediction harder, as there is more inherent flexibility. The lack of interest is due to the historical misunderstanding of the large role of RNA in biology. The main interest in structural prediction was focused on the presumed functional machinery of the cell, composed of proteins.

Our initial computational interest was mainly directed toward developing tools that could deal with poor structures. These would be useful for cleanup of poor geometry from various different sources, crystallography, *de novo* generation, homology modeling, NMR and others. The volume of structures that need processing also argued for developing algorithms that are rapid.

Prediction of secondary structure is arguably easier for RNA than for protein[9]. There are strong base paring rules, and therefore, more completely determined thermodynamics, that allow better prediction to be made for the secondary structure[10].

This gives a foothold into tertiary prediction; the current theory of folding progression in nucleic acids suggests that most of the secondary structures forms first, then the structure searches for its final 3D structure[11]. This is in contrast to protein structure that starts with an initial amorphous collapse, then a search for a final conformation.
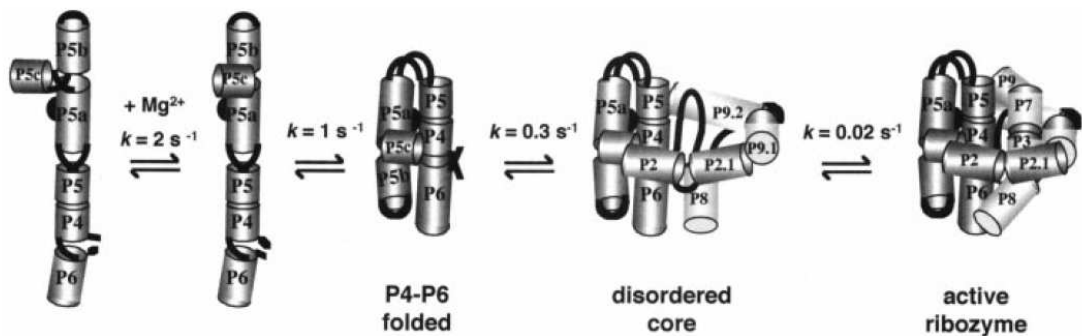
Figure 1: Folding pathway of the *Tetrahymena* ribozyme as determined by synchrotron hydroxyl radical footprinting[12]. This experiment shows secondary structure forming before tertiary structure.

Secondary structure prediction gives us a starting point for structural prediction[12]. Our goal was to develop better predictive tools, but to use the secondary structural information effectively for tertiary prediction we needed better understanding of the structural rules in RNA and DNA. This would give insight into how to search effectively. Since full quantum mechanical treatment of large molecules, like nucleic acids, is currently impossible, we decided to consider nucleic acids as geometrical constructs with classical mechanics.

2.2 Torsion Angle Preference

So what is in a structure? A structure is represented as the x, y, z, Cartesian coordinates of all the atoms in a molecule. Alternatively, if the bond lengths and bond angles are held constant, the structure can be represented as the torsion or dihedral angles

for all the single bonds in the structure. On the timescales shorter than that of bond making and breaking, chemical bonds can be approximated as stiff rods; so we took for a starting point the dihedral angles in the backbone of the nucleic acid. Dihedral angles are defined as the angle between two planes, in the context of a molecule four atoms (or lone pairs) about a central bond can sufficiently define these two planes[13].
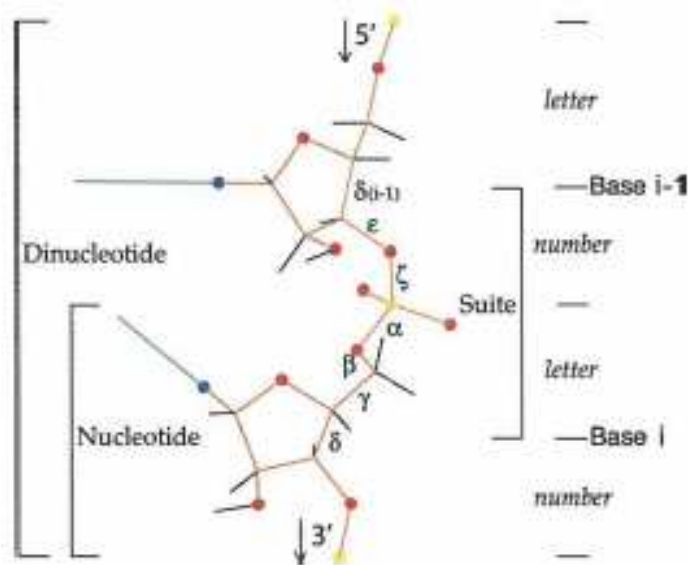
Figure 2: The dihedral angles present in the backbone of RNA[14]

These torsion angles act as the joints about which nucleic acids can flex to fold into stable structures. This restriction to dihedral angle flexibility reduces the structural search space allowed to a molecule dramatically. For example, each residue in RNA has about thirty three atoms on average. This means that the number of coordinate parameters is approximately $3 * 30 * N$, where N is the number of nucleotides. In dihedral space, the number of rotatable single bonds is $7 * N$ (six backbone and one side chain dihedral). We therefore explored the energetic landscape of allowed geometry by the dihedral angles in natural structures, and computationally the space of possible conformations. The

computational work was necessary due to the fact that the native structures are generally crystal structures, a snapshot in time of a dynamic structure, frozen in a crystal by neighboring molecules, perhaps with non-biological additives, and held at cryogenic temperatures (typically 77 K) to reduce radiation damage and slow molecular motion; all done to improve the resolution of x-ray methods, or allow them to be accomplished in the first place. These experimental constraints of the crystallographic process raise the question of the actual conformation in functional, biological conditions. Also, the vast majority of structure determinations are funded for biologically functional molecules; functional biomolecules are the end point of the folding process, which does not properly describe the conformational process required to reach the fully folded state[15].

The first part of our analysis of RNA structure was an assessment of conformations observed in functional nucleic acid structures. We hoped to reduce the size of the conformation space that our later design methods would have to search. Perhaps we could also find some underlying principal guiding nucleic acid assembly and function. We also hoped to discover smaller structural units that would allow a modular building approach. This was started with a survey of the angles present in the 30S subunit of the ribosome. We chose the crystal structure of the *Thermus thermophilus* 30S ribosomal subunit (1J5E PDB identification number). For all dihedral angles a bin size of -0.499 to +0.500 degrees was used. We found that chi, zeta, epsilon, beta, and alpha have one area of dihedral angle space that is preferred, but that delta and gamma have two. This is in broad agreement with previous work [16].

Alpha Histogram
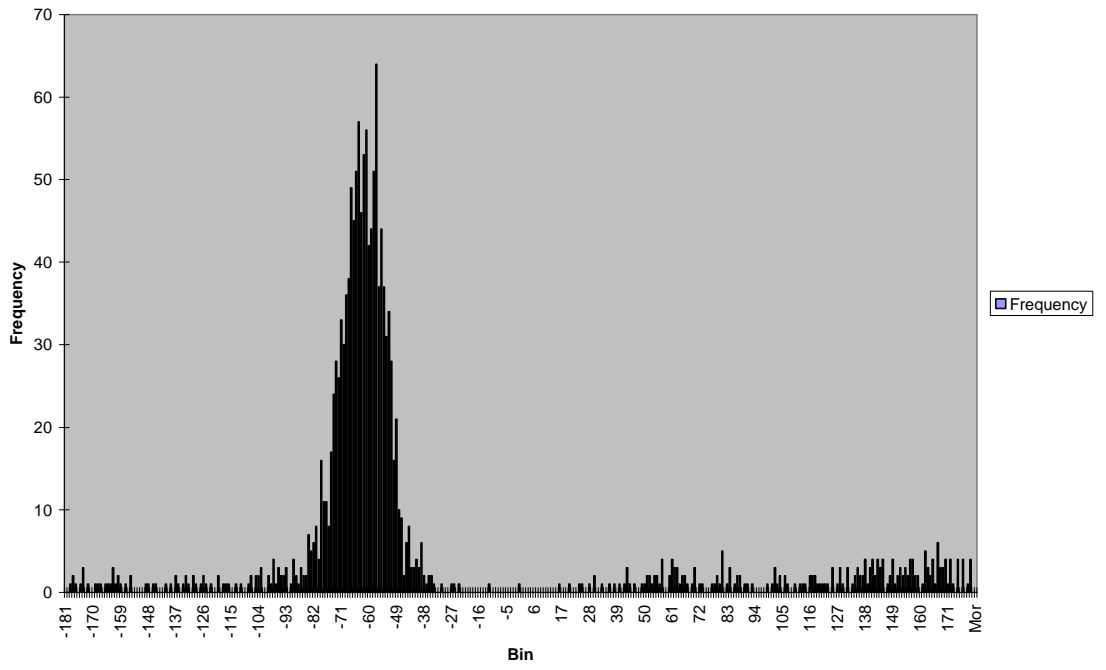
Beta Histogram
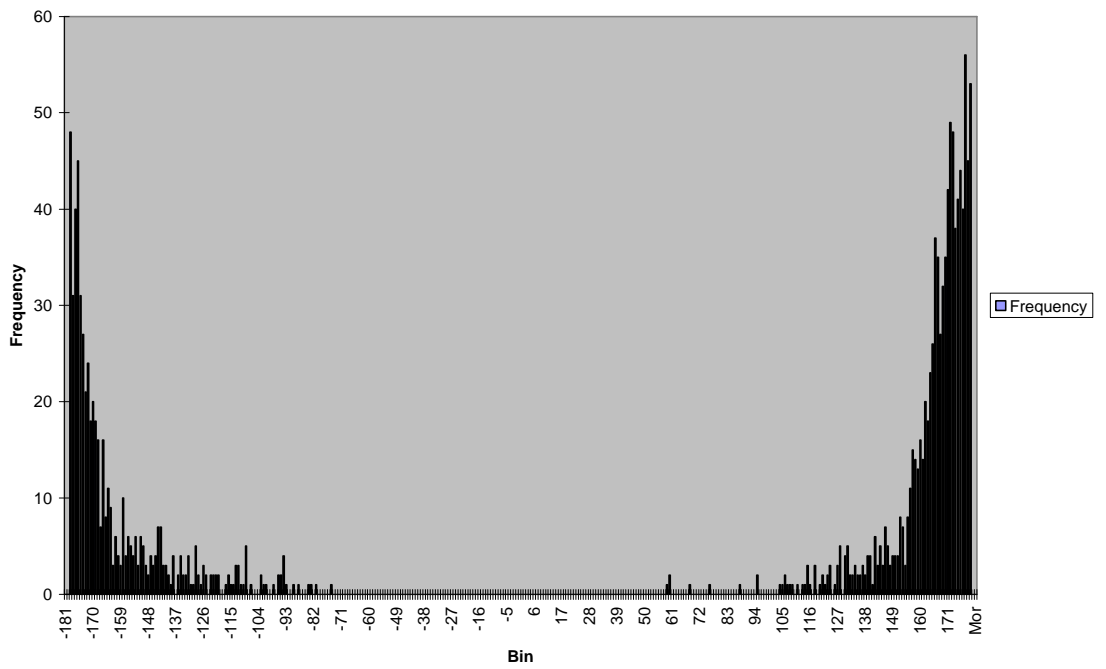
Gamma Histogram
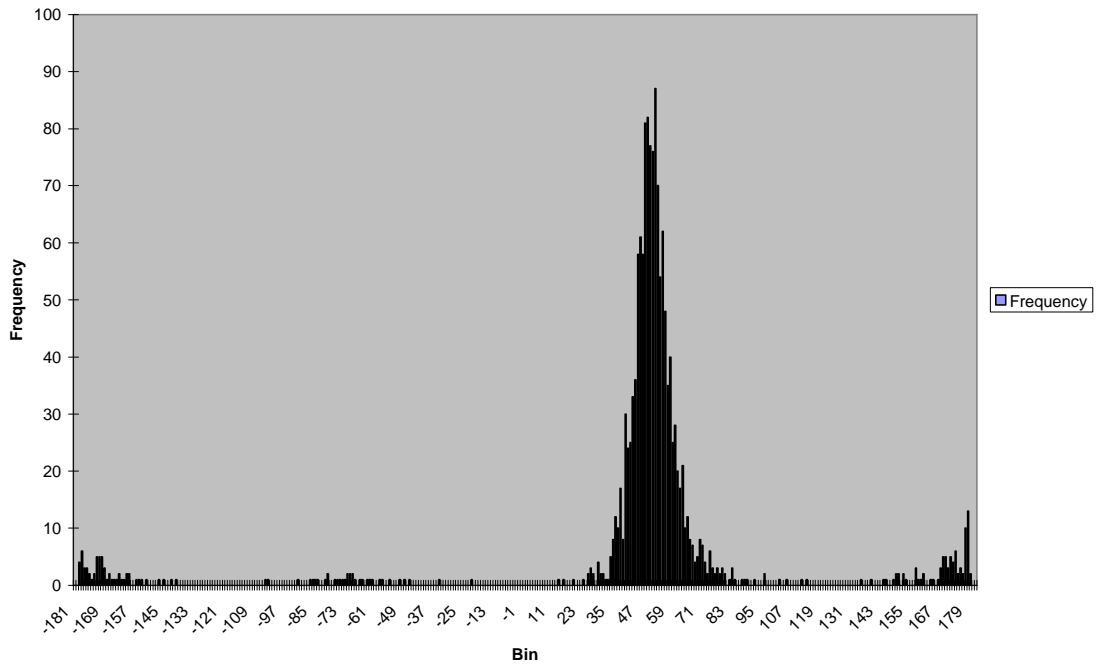
Delta Histogram
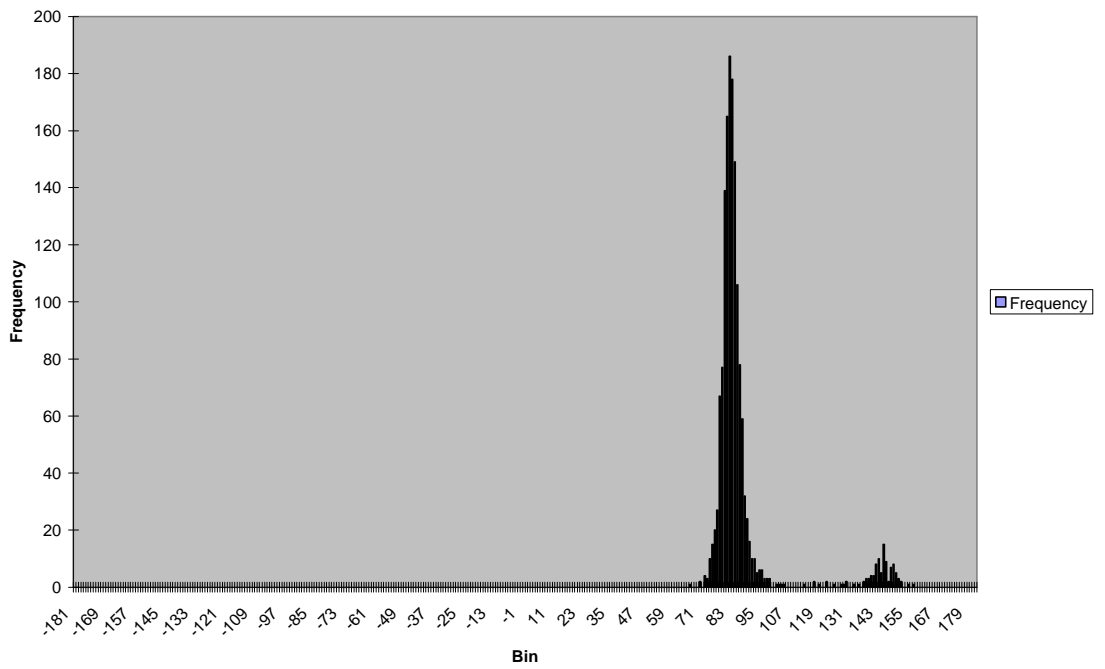
Epsilon Histogram
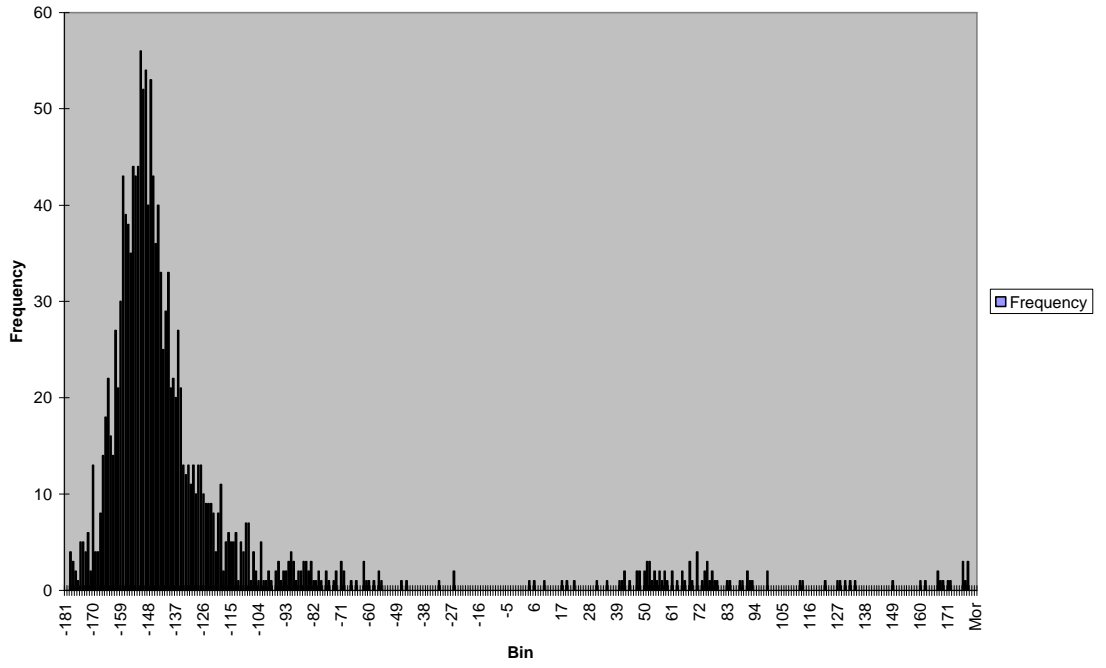
Zeta Histogram
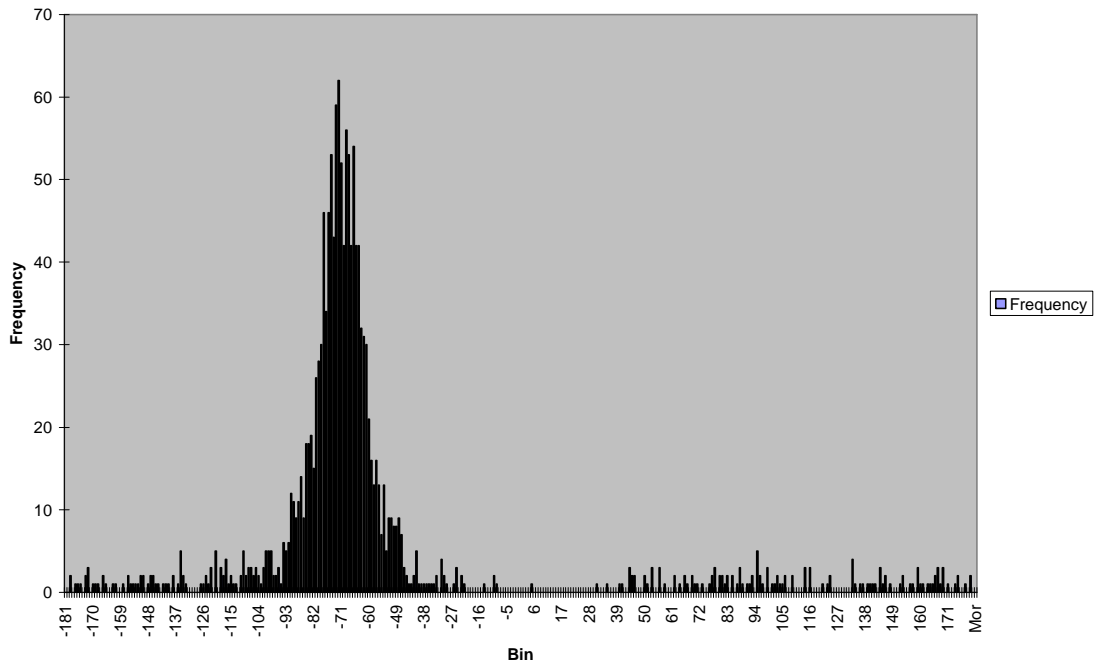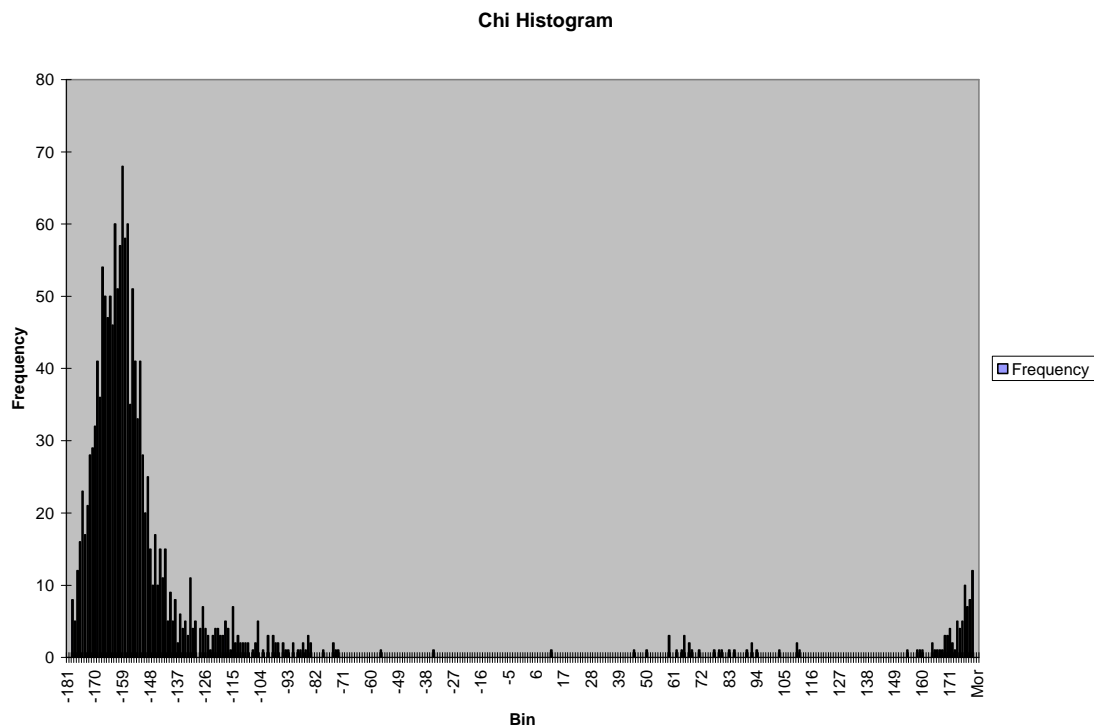
Chi Histogram

Figure 3: Dihedral angle histograms calculated from the 1J5E crystal structure
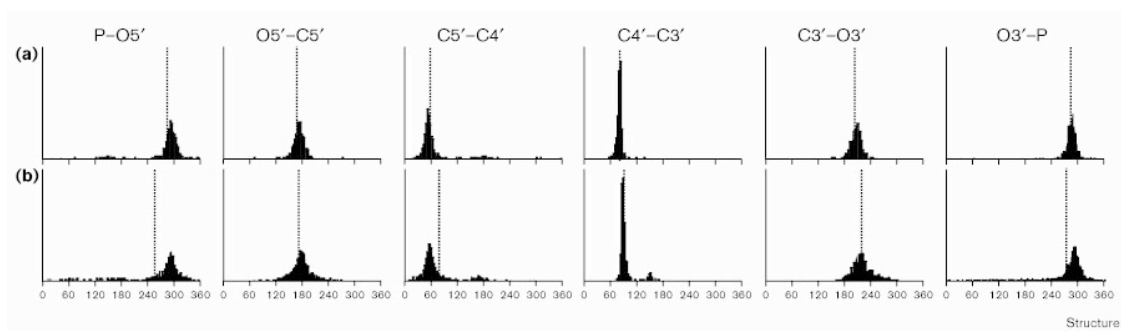
Figure 4: Dihedral angle histograms calculated by Westhof and Fritsch[17]. The histograms labeled with A are constructed from helical fragments, those labeled B are constructed from large RNA structures. All structures were from the Nucleic Acid Database circa 2000. To be comparable to our values, values higher than 180 degrees map to negative 180 and then down to zero degrees (*i.*e. 181 here maps to –181 and 359 maps to -1)

The width of the angular space preferred depends upon the angle and the trend is

chi, epsilon, beta, zeta, alpha, gamma, delta, from widest distribution to smallest. The

angular freedom shown here roughly mirrors the number of atoms attached to atoms in

question. For example, epsilon and beta are the dihedral angles that involve oxygen atoms from the phosphate that have no other groups attached other than those continue the thread of the backbone. Zeta and alpha also include the same phosphate oxygens that have no other substituents, they are more restricted in flexibility than epsilon and beta since they have to contend with the oxygens (*i.e.* O1P and O2P) attached to the phosphorus. The two preferred delta angle positions reflect the two different low energy sugar pucker conformations possible for the five membered ring of the ribose[16b]. The conformational rigidity of the ring also is reflected in the narrowness of the distribution of the preferred angles. The chi angle is known to prefer two conformations, *syn* and *anti*. The low population of the syn conformation is partly due to the analysis being done on a highly structured RNA; normal base paring favors the anti conformation[16c, 16e]. Also, pyrimidines (*i.e.* C, U and T) are only rarely in the *syn* conformation due to the steric clash between O2 and H3'. The conformations of the alpha and zeta dihedral angles have three preferred minima, with *gauche⁻* the dominant observed dihedral, particularly in base paired regions. These angles are probably restricted by the phosphate that is part of the atoms that make up these dihedral angles. Epsilon and beta have more flexibility, as they are one bond away from the confining sugar. Gamma has two preferred geometries *gauche⁺* and *anti*, and one rare geometry *gauche⁻*, and is restricted in flexibility by the presence of hydrogens on the ribose ring. The question is then, with few usually allowed angles in a RNA structure, how is the large amount of geometrical diversity obtained? One way is that small changes in an angle over a long distance can add up to a large distance in space. This is known as the lever arm effect[18]. The other mechanism for structural diversity is that infrequently found rare angles can add structural diversity. The

demonstrated preferences allowed us to proceed with some confidence in reducing the space to be searched in the first approximation to the preferred angles, and a sampling of a few rare angles (M. Kockhal and J. SantaLucia, Jr. unpublished results).

2.3 Force Field Validation

The choice of forcefields for energy calculation is important for classical molecular mechanics studies. Care in selection and validation of forcefields is necessary since we planned to use the calculated energy for a particular conformation as a metric to guide a structure to a folded state. An assumption is made that the lowest energy structure is the native, biologically relevant, form of the molecule[19]. Thus, if energy can be accurately calculated, it can guide a folding algorithm to the native, folded form of the molecule[20]. To guide our simulations, we needed a metric of structure quality. A metric is needed by the computer algorithm to deduce if a trial structural change is better or worse than the original structure. This metric will act as a guide to fold a structure. Usually this metric is some sort of energetic calculation for physics based approaches, or combination of "knowledge based" approaches. The assumption is that biologically relevant structures use thermodynamic minimization of energy to fold to a functioning structure. Since nature performs free energy minimization, a computer can mimic this process with a calculated energy. If the functions used to calculate *in silico* energies are close enough to real world thermodynamics, then the molecule will be guided to the natural structure. One obvious problem with this approach is that energetic barriers can prevent algorithms from finding the energy minimum conformation. Interestingly, this also seems to occur naturally in RNA; some structures can be kinetically trapped during transcription or translation [21]. This problem may then be a real aspect of nucleic acid folding.

A force field for energy calculation usually has the general form of a sum of terms representing the various physical forces and principals that are thought to (semi-classically) drive molecular and atomic association. These terms are usually partitioned into a harmonic bonding force (this represents the force of a covalent chemical bond in a distance dependant fashion), bond bending (this represents preferred bond angles defined by two adjacent chemical bonds), a dihedral angle component (due to atomic packing around most dihedral angles, there are preferred conformations in molecular dihedral angles, this term accounts for this observation) , electrostatics (this term is the classic columbic attraction or repulsion of charged bodies), and a Van der Waals term (this component represents induced dipole-dipole interactions, and Pauli repulsion when atoms are too close). The forcefield we used was slightly modified from this general description, as we added a hydrogen bonding term, as these interactions are very important to nucleic acid structure, and are not adequately represented[22].

$$E_{total} = \sum_{P-O3'bonds} K_r (r - r_{eq})^2 + \sum_{P-O3'angles} K_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + C_{ij} \right]$$

$$+ \sum_{i<j} \left[\frac{D_{ij}}{R_{ij}^{10}} - \frac{E_{ij}}{R_{ij}^6} + F_{ij} \right] \times (angle\_terms) + \sum_{i<j} \frac{q_i q_j}{4 * pi * \varepsilon_o * D * R_{ij}}$$

Figure 5: The Energy functions used in our simulations, referred to from here on as NA-FF (Nucleic Acid Force Field)

Different portions of the potential function were screened for utility in guiding our folding and ranking work. Two common functions in use are the Buckingham potential[23] and the Lennard-Jones potential[24]. Both are to model the pair wise interaction terms that reflect the repulsion due to electronic orbital overlap when atoms become too close, and

the attraction due to Van der Waals forces (induced dipole dipole interactions). There are a few arguments for using either potential. The Lennard-Jones potential has steeper slopes than the Buckingham potential, goes to infinity as atoms become close, and is easier to calculate than the Buckingham potential. The Buckingham potential is "softer" than the Lennard-Jones potential, has a better basis in real physical phenomena (the exponential term mirrors the Pauli Exclusion Principle more closely), but is harder to calculate and can have relatively low energy values at very close atomic separation. The way that one decides what to use in an empirical force field is to see what functions and parameters work best for the purpose to which it is to be put. To see which potential would be more useful for the nucleic acid folding problem, we chose the small well defined RNA crystal structure as our testing goal (the sarcin-ricin domain from *e. coli,* PDB id 483D. At the time of the work it was the RNA structure with the best resolution, which still displayed a reasonable variety of structural motifs; see the Dickerson dodecamer for a higher resolution structure that shows less structural information, 1DPN). To generate a test set of structures we adopted two strategies. The first was to randomly change the dihedrals in the structure with a certain amount of allowed angular change. The second was to alter the angles with a gaussian distribution. The first strategy would produce bad, and often non-physical structures, the second would mirror the error in experiment better, and produce higher quality decoy structure. We then ranked these test sets by the calculated energy and RMSD to the native crystal structure. We found that the energy function did not correspond well to coordinate closeness unless a round of energy minimization was done first to clean up gross structural problems, which made the

energy essentially random. After this step, both energy functions seemed to correspond to

closeness to the native structure.

$$E(r) = be^{-r/\rho} - \mu r^{-6}$$

Figure 6: The Buckingham potential

$$E(r) = \lambda r^{-12} - \mu r^{-6}$$

Figure 7: The Lennard-Jones potential
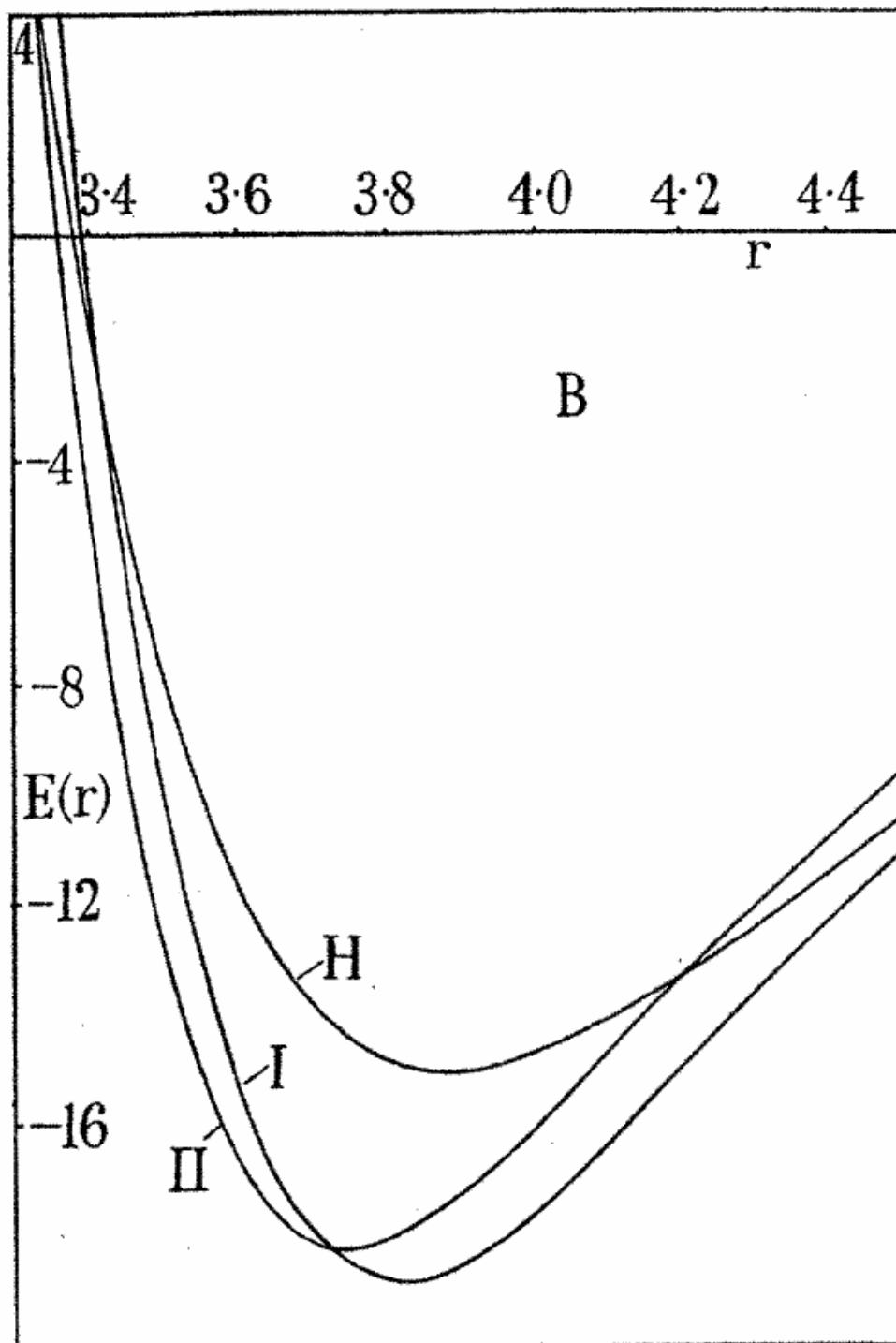
Figure 8: Graph of the Lennard-Jones and Buckingham potential functions for Argon atoms in the gaseous state. The Lennard-Jones potential is curve II, the Buckingham potential is curve I (curve H is the Herzfeld interaction derived Herzfeld and Goeppert-Mayer's exact equation of state for argon[25]). R is the distance between two Neon atoms in angstroms, E is the energy of the interaction in ergs X 10^{-15} (X10^{-22} joules, X 6.24*10^{-4} electron volts) The functions have their parameters fitted to replicate measured physical data.

We decided to determine if our energy function could rank structures according to their distance from a native crystal structure. To generate structures that would be near a known structure we altered the dihedral angles of a well known structure, subject to a Gaussian distribution centered on the original angular values. This allowed us to generate "decoy" structures with a known folded structure. This also allowed us to have a secondary metric to judge the utility of our energy function, the root mean square deviation (RMSD) of the decoy structure from the original crystal structure. The amount of angular deviation introduced into the structure only corresponded weakly with the calculated energy and the RMSD. In terms of the energy, this is most likely to be caused by a crashing effect of one part of the structure into another. Only a small change in dihedral angle can create a non physical structure, where atoms are overlapped. A large energetic penalty for non physical structures is incorporated into the energy function to select against them, so these sort of nonphysical structures cloud the selectivity of the energy function. This clouding makes it is difficult to tell the difference between a structure that is geometrically close to the native structure, but has some non physical clashes, a completely unfolded structure, and one with many small problems. The magnitude of the calculated energy is often poorly correlated with the 'quality' of the structure. The solution to this problem was the addition of energy minimization before the energetic and Gaussian assessment. This step cleaned up these nonphysical structures to the point that the energy function, RMSD and the amount of angular deviation introduced generally trended in the same direction.
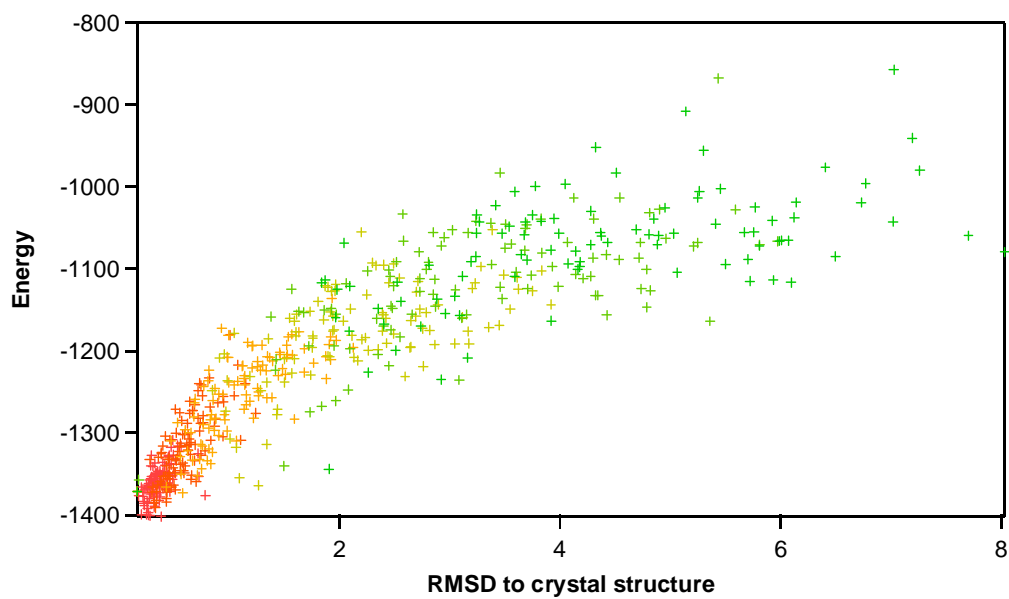
Figure 9 : Scatter plot of the NA-FF calculated energy versus the RMSD to the native crystal structure for gaussian generated decoy structures. The different colours correspond to the width of the standard deviation used to generate the structures.
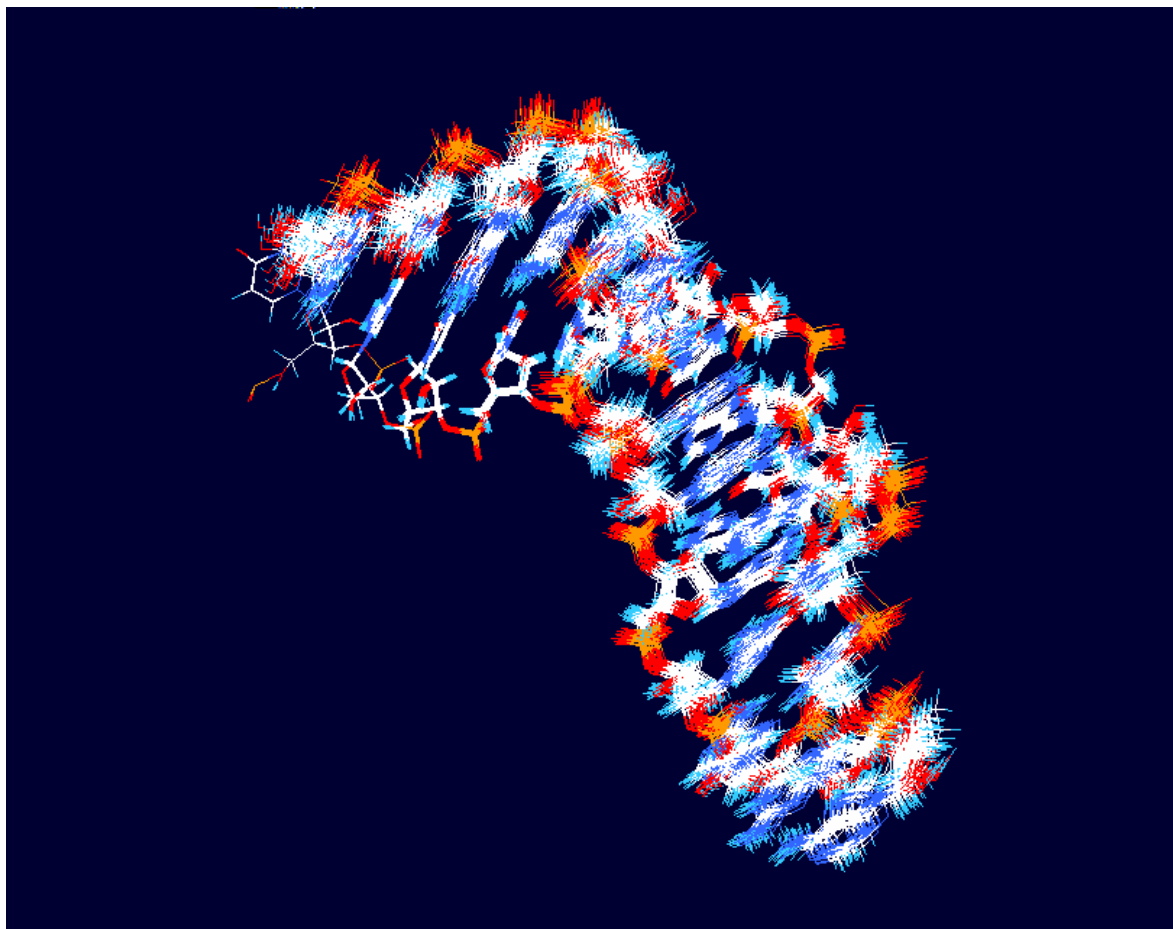
Figure 10: Picture of the ensemble of structures generated with gaussian angle sampling. These structures were generated using a standard deviation of half of a degree.
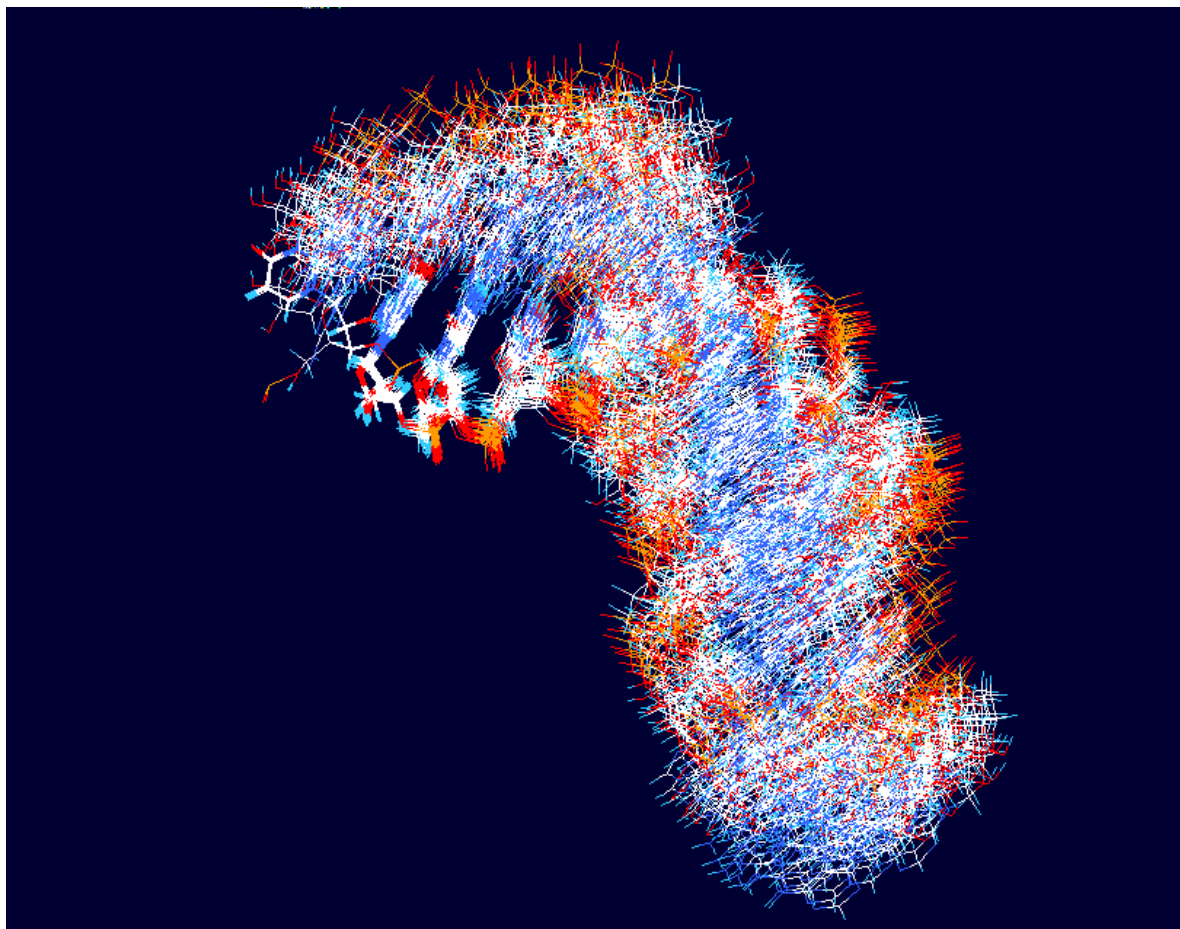
Figure 11: Picture of the ensemble of structures generated with gaussian angle sampling. These structures were generated using a standard deviation of one degree.
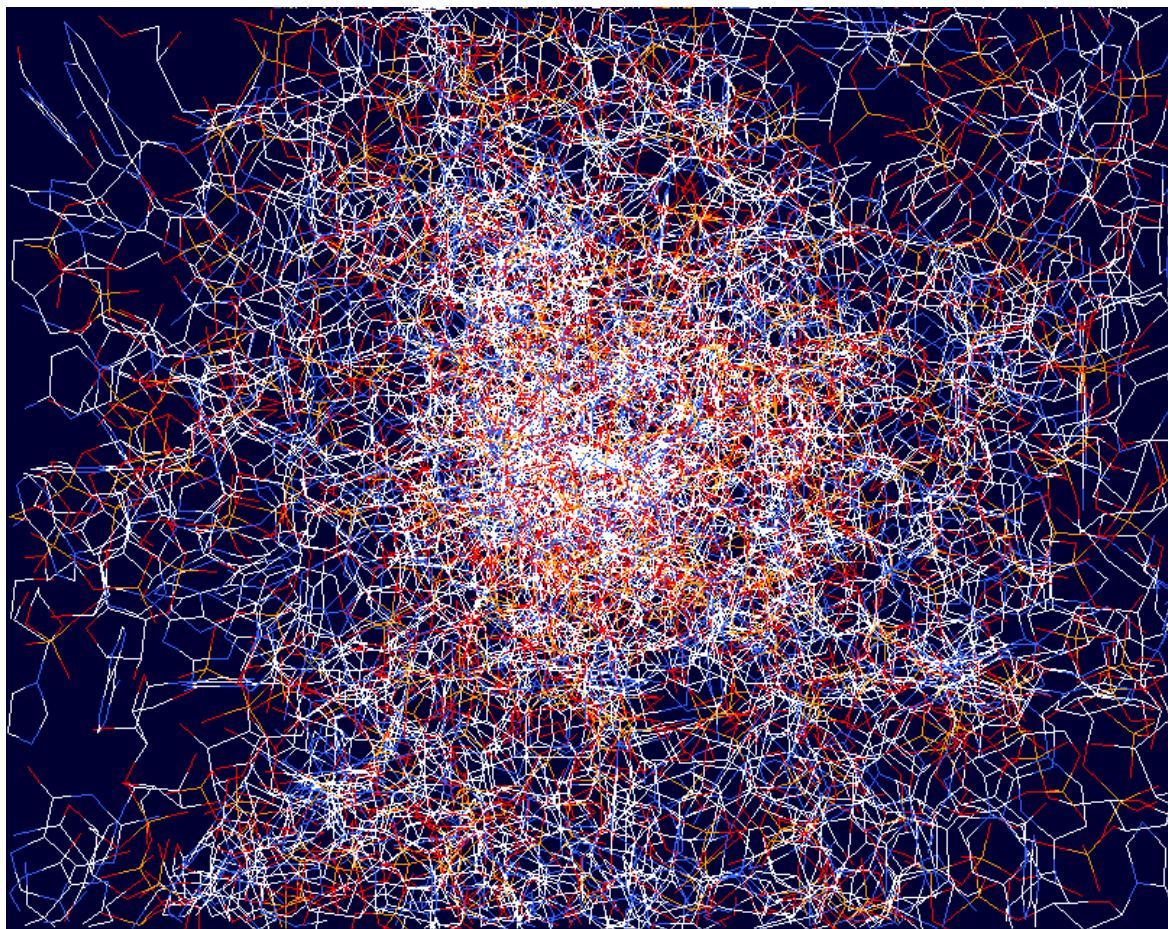
Figure 12: Picture of the ensemble of structures generated with gaussian angle sampling. These structures were generated using random values for their dihedral angles.

2.4 Comparison of NA-FF with AMBER

To further assess the use and application of our forcefield we decided to see if it could select correct, native structures from a larger set of incorrect decoy structures (*i.e.* to test the discrimination of the forcefield). The set of correct structures was twenty four different 8-mer RNA hairpins taken from various different crystal structures. Appendix 1 shows the sequences and origins of the hairpins used. Decoy structures for the correct structures were created by "threading" the sequence associated with a particular hairpin, through all the other hairpins in the sample. In practice, this

meant keeping the geometry of a particular structure, but replacing the native bases with those from the other structures (the chi dihedral angle of the base was preserved in the replacement process). This process generated 576 (twenty four structural motifs times the twenty four sequences that were native to each motif) structures in the set, from which the goal was to pick the 24 correct structures (the structures with the native nucleotide sequence in the native structural motif) by ranking these structures as the lowest in energy. As shown earlier, DSTA energy minimization was performed on the structures to reduce the number of gross structural problems. There were some problems with this approach. One was that the structures were in the context of larger macromolecules. Their isolated energy (or conformation) then may not correspond to the energy in the context of the whole native biomolecule. This is because long range and tertiary interactions can stabilize different folds than would be seen in isolation. Another problem may that some of the hairpins may be in extended conformations that show little selectivity for a particular sequence, as the bases are not in positions to interact with each other regardless of sequence. *In vivo,* an analogous process is mutational tolerance. Nevertheless, we had planned to use our energy function in homology modeling, where we envisioned a process much like this to be used to replace mutated portions of a known molecule. The same test set was also ranked with the AMBER forcefield [26], to see if a modern molecular mechanical forcefield would do a better job of discriminating between native and non native structures. As can be seen from figures thirteen and fourteen, the NA-FF (Nucleic Acids Force Field) does a better job of discriminating between decoy and native structures. AMBER, in fact does no better than random chance in picking the correct structure and sequence.
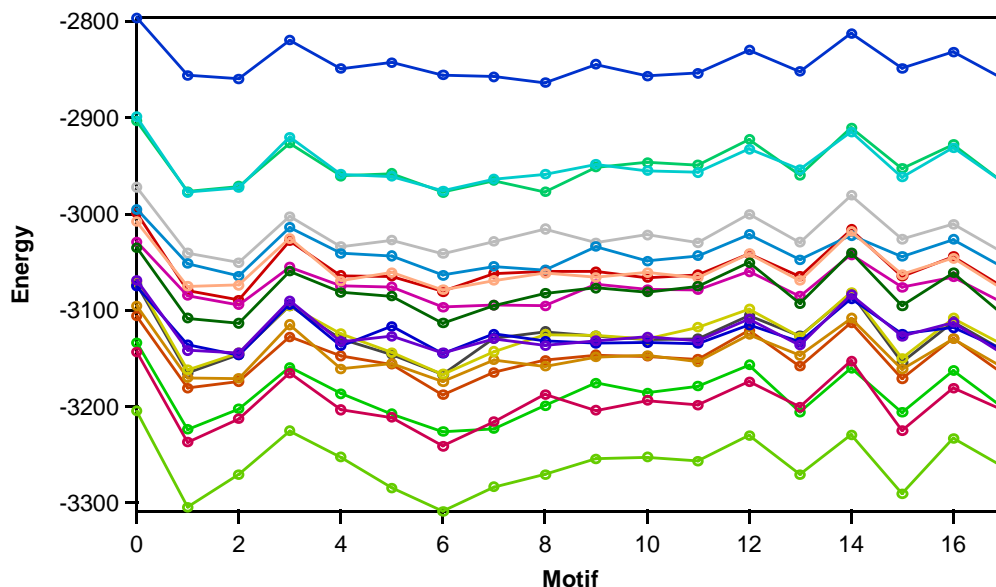
Figure 13: Amber prefers particular sequences (and motifs), no matter which structural motifs they are threaded through

	AMBER	NA-FF
correct structure rank average	10.81	9.88
number correct	2	3
% correct	7.69	11.54
correct structure in the top quartile	11	13
% top quartile	42.31	50

Figure 14:Ranking summary for hairpins of length 8. The data was generated by taking 18 different hairpins of length of eight from RNA crystal structures, and threaded every sequence from the 18 into every structure, giving a set of 306 decoy structures with 18 correct structures. The resultant structures had their respective energy calculated with each forcefield and these were used to generate the rankings.

From figure thirteen, the AMBER forcefield appears to have a preference for particular motifs and sequences regardless of the context. AMBER's poor discrimination is probably due to the fact that the forcefield is designed to deal with physically plausible

structures, and not with the physically non-plausible structures that are often generated as decoy structures. The most obvious preference is that AMBER consistently ranks sequences rich in G's and A's as higher in energy. We hypothesize that the preference for particular sequences and structures that AMBER displays is probably due to poorer positioning of purines when they are in a crowded position. If the motif has a lot of space free space around the nucleobases AMBER rates it as low in energy regardless of sequence as there are fewer positions where the purines can be crowded. Conversely, if a sequence is rich in G's and A's AMBER rates it as higher in energy than the other sequences, as there is a good likelihood that one of the purines will be in a bad position. AMBER may perform poorly in this comparison as its energy minimization algorithm is taking steps with all atoms, and thus we implemented DSTA. This torsion angle sampling is more efficient computationally than simulating the trajectories of all atoms involved in the biomolecule; it may not get caught in some types of minima that a steepest descent procedure can be mislead into.

2.5 Conclusion

What was the result of our testing? We found for our purpose that the Van der Waals portion of the energy function made little difference in discriminatory power for native structures. We found that the energy function did correspond to closeness to a native structure (measured by RMSD) within a window of distance to the native structure. Also, our energy function appears to do a better job than a standard molecular dynamics package (AMBER) in finding a native structure obscured by non-native sequences. This gave us confidence to use the forcefield for guidance in the package of

programs that became RNA-123 and when licensed and much improved by DNA Software Inc., became Nucleic Acid CAD^{TM} (NA-CAD^{TM}), a commercial product.

3. DNA Nanostructural Design and Synthesis

3.1 Introduction

With a basic forcefield for nucleic acid simulation developed, we continued on to the design and synthesis of nanostructures. The current state of the field involves tedious manual design of self assembling structures or truncated models of known thermodynamics. These approximations become less and less possible with the increasing complexity of the system. Also, because of the inefficiencies of design, a complicated structure can mean many rounds of purification and characterization. If a simple method for designing structures that would assemble in one pot in one step could be found, it would be very useful to the field. The SantaLucia lab has a history of successful computational thermodynamic prediction of nucleic acid association, and based on this previous work we developed prediction software that may allow the design of more complicated structures. We also experimented with light controlled "tack welding" of assembly or disassembly of a DNA structure which could add more precise controllability to the system and perhaps mechanical functionality.

Nanotechnology promises to deliver new 'smart' materials with tunable properties, efficient catalysis, novel drug delivery mechanisms, and nanomachines [27]. The control of materials of the nanoscale can be accomplished from two directions, bottom up or top down. The top down approach is to start with large objects and shrink them down in some way. A good example of this is the fabrication of integrated circuits. A large design is made into a nanoscale design with photo reduction. A large pattern is made into a nanoscale structure by optically focusing the design of the pattern to a smaller size on a photoactive substance. The area of the photoactive substance that has not been

illuminated by the pattern can be washed away with solvents leaving a replica of the larger pattern at a smaller size. The other direction for constructing a nanoscale structure is from the bottom up. This means that one starts with atoms and assembles them into large (on the atomic scale) structures. As chemists, we have good knowledge of the atomic level assembly of matter into small structures (in the angstrom size regime); however, making larger nanoscale structures is still challenging. With our experience in the chemistry of DNA folding and hybridization giving us atomic control, our choice of a bottom up approach is obvious.

3.2 Computational Design of a Nanostructural DNA Triangle

DNA is a good candidate as a nanotech building block because of its stability, the ability to encode into the molecule specific associations via the Watson-Crick base pairing rules, the multitude of molecular biology tools for specific molecular manipulations, and the ability to reliably replicate and synthesize a molecule or large quantities. DNA was selected over RNA for our experiments to reduce the initial experimental complexity (DNA is more chemically stable than RNA, can be more cheaply synthesized, and more enzymes are available for processing.). DNA in biological systems also is known to assemble into nanoscale structures with exquisite control of its conformation and function (*e.g.*, packing of DNA into viral nuclei or the chromatin structures in the nucleus of a eukaryotic cell).

The task of specifying the sequence for a DNA strand so that it associates in a specific manner with other strands has generally been done by hand in the past. This approach has a considerable drawback. It is very difficult for a human to go through a long sequence, or many shorter sequences, and eliminate all the possible competing

interactions that a particular sequence, or sequences, may have. These competing interactions seriously reduce the yields for assembly and require many purification and assembly steps [28]. More recent computational algorithms have typically used a truncated interpretation of the known thermodynamic complexity in nucleic acids. While the computational methods have met with much success, we believe the same problems as with manual design will be found for more complicated structures; that the computational algorithms will not scale well, or be robust, due to incorrect thermodynamic modeling. To alleviate this problem, the SantaLucia lab created a computer algorithm to automatically design a set of DNA sequences that assemble in a specified manner. Using the thermodynamic nearest neighbor model, the rules of an expert in the field, positive and negative design features, and a evolutionary search strategy, the program mutates a starting set of sequences to give a final set, that assemble in a programmed manner (see figure fifteen for the block diagram of the program). The program redesigns any sequences created that form undesired structures, individually or for any pair of sequences in a group. Using the program, we designed an equilateral DNA triangle; approximately 16 nanometers on a side, comprised of nine different strands that should assembly correctly in one pot, in one step (see figure sixteen for the predicted secondary structure of the triangle, and figure seventeen for the predicted 3D structure of the triangle). We attempted to use an enzyme called T4 DNA ligase to seal the junctions between the strands, giving a stable structure that is resistant to disassembly. Ligase will catalyze the formation of a covalent bond between the 5' phosphate of one strand and the 3' hydroxyl of another strand, if they are held next to each other *via* base pairing by a strand that is complementary to the strands to be ligated. For example, planned in our

design, if the ligation is successful, sequence 1 and sequence 2 (a 39mer and a 41mer) will be joined into one 80 base long piece of DNA. Sequences 1 and 2 are held together for the ligation by sequence 7 (see figure 16 for the specific base pairing). If all the designed ligations are successful there will be 3 different 80 nucleotide pieces (sequence 1 joined to sequence 2, sequence 3 joined to sequence 4, sequence 5 joined to sequence 9) and one circular 117 nucleotide central piece (sequence 7 joined to sequence 6 and to sequence 8). The creation of these larger pieces can be used, in part, to show the correct assembly of our designed structure. The initial triangle was to be used a proof of principle and a building block for more complicated structures.

Figure 15: Block diagram for the DNA design program
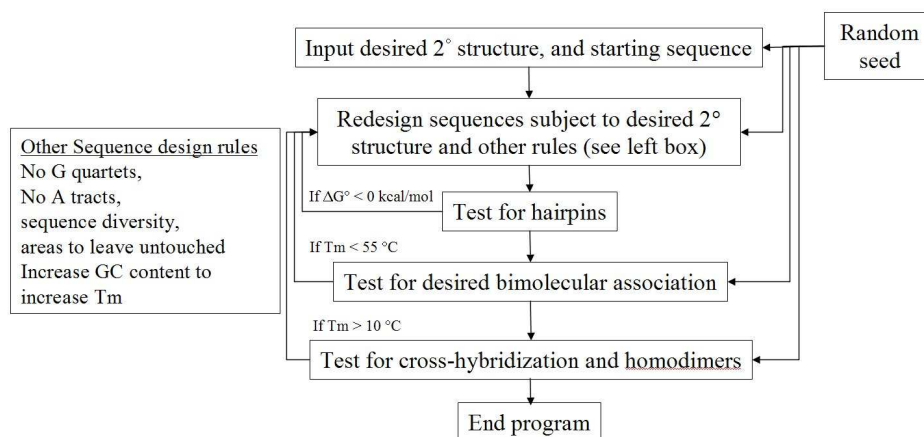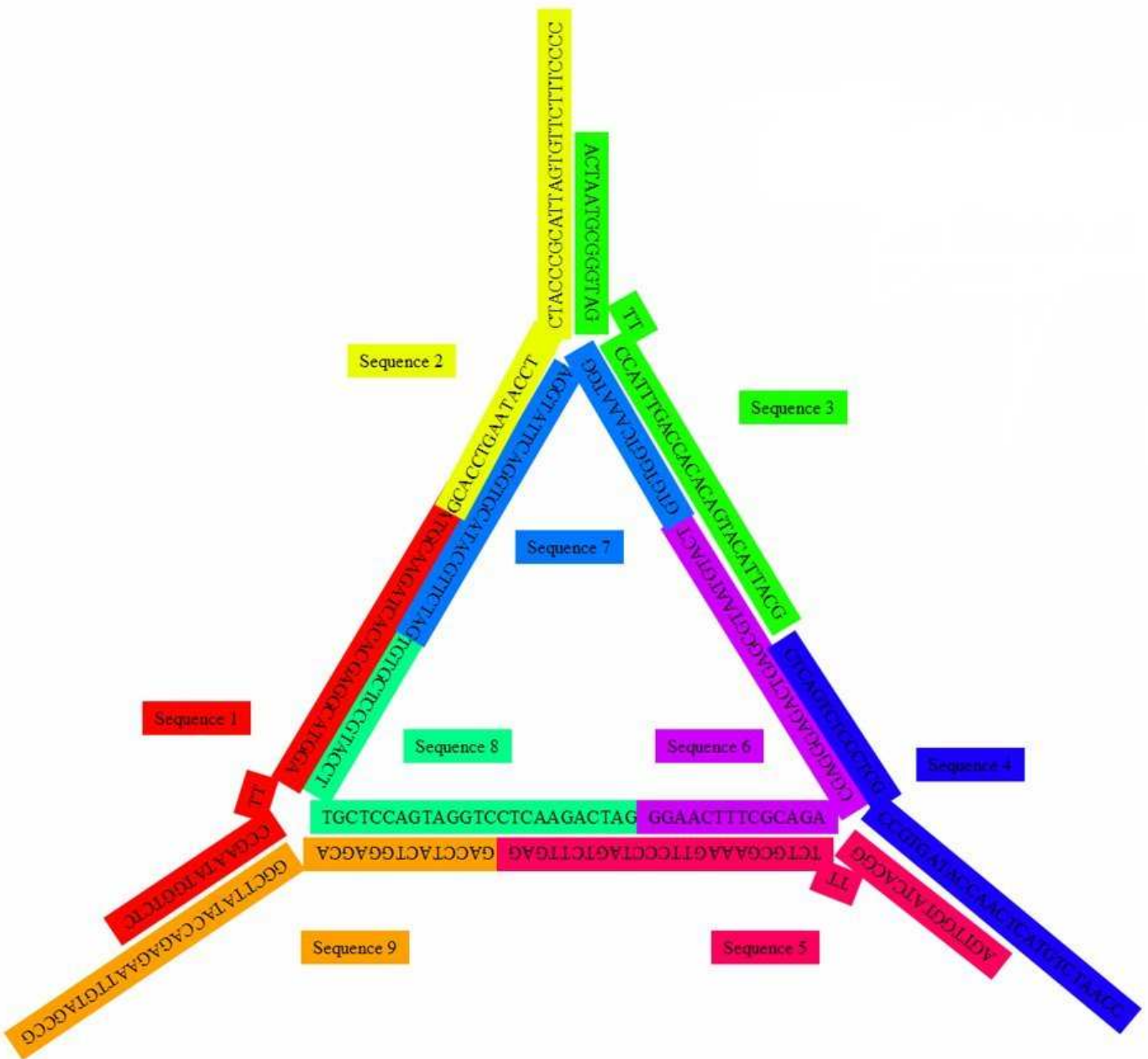
Figure 16: Secondary structure of the initial triangle design
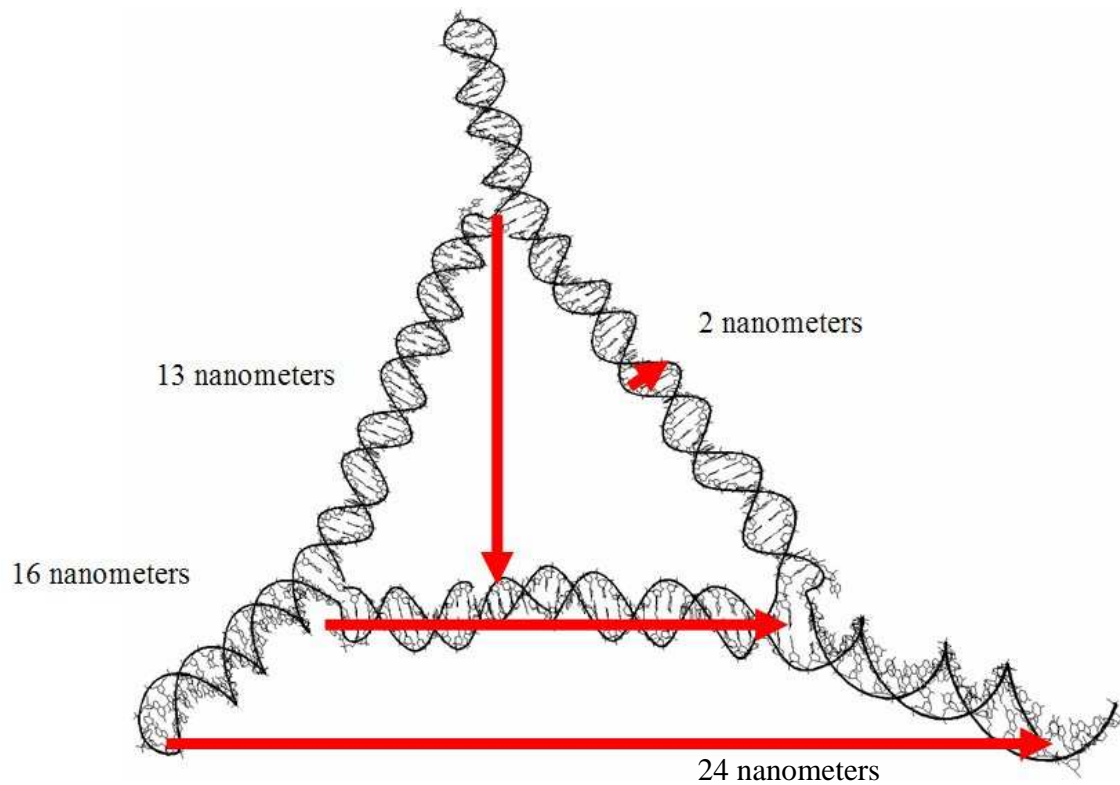
Figure 17: Predicted 3D structure of the initial DNA design

3.3 Synthesis and Characterization of the Triangle

3.3.1 Ligation

The designed strands were chemically synthesized by Sigma Genosys, and ligation was attempted with T4 ligase and Taq ligase using various conditions (see figures 18, 19 and 20). Two different ligases were used to determine which was more effective in the assembly process. T4 DNA ligase is very active, but somewhat nonspecific and unstable at the temperatures simulated in our design algorithm(*i.e.* 55°C)[29]. Taq ligase is very stable and specific[30] at temperatures up to 85°C. We used polyacrylamide gel electrophoresis to assess the ligation reactions. Native gels should

preserve any assembled triangles; denaturing gels should break the triangle into single
stranded pieces.
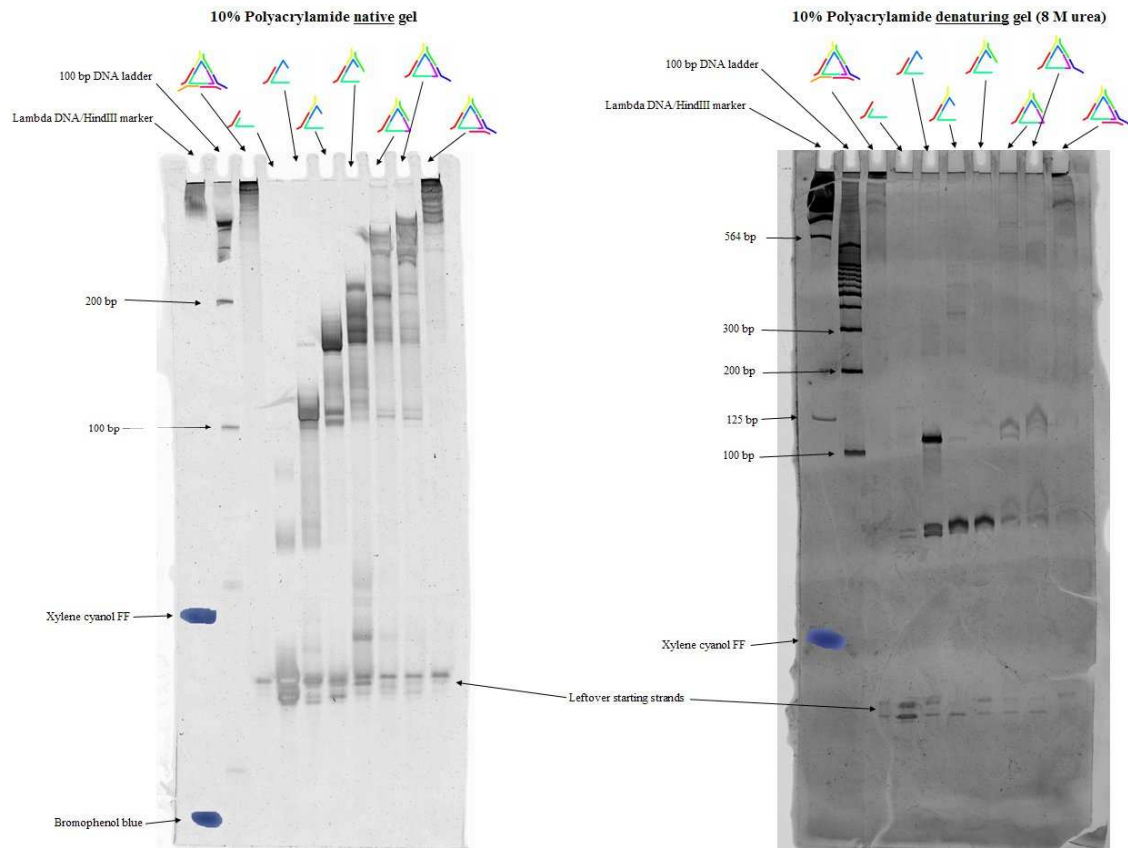
Figure 18: Native and Denaturing Polyacrylamide Gels showing T4 ligation Studies. Note the absence of a clean band of triangular product, which should be near 117 bp. Also see that ligation has occurred, indicated by the presence of large pieces of DNA (greater in size than the starting 40mers) on the denaturing gel.
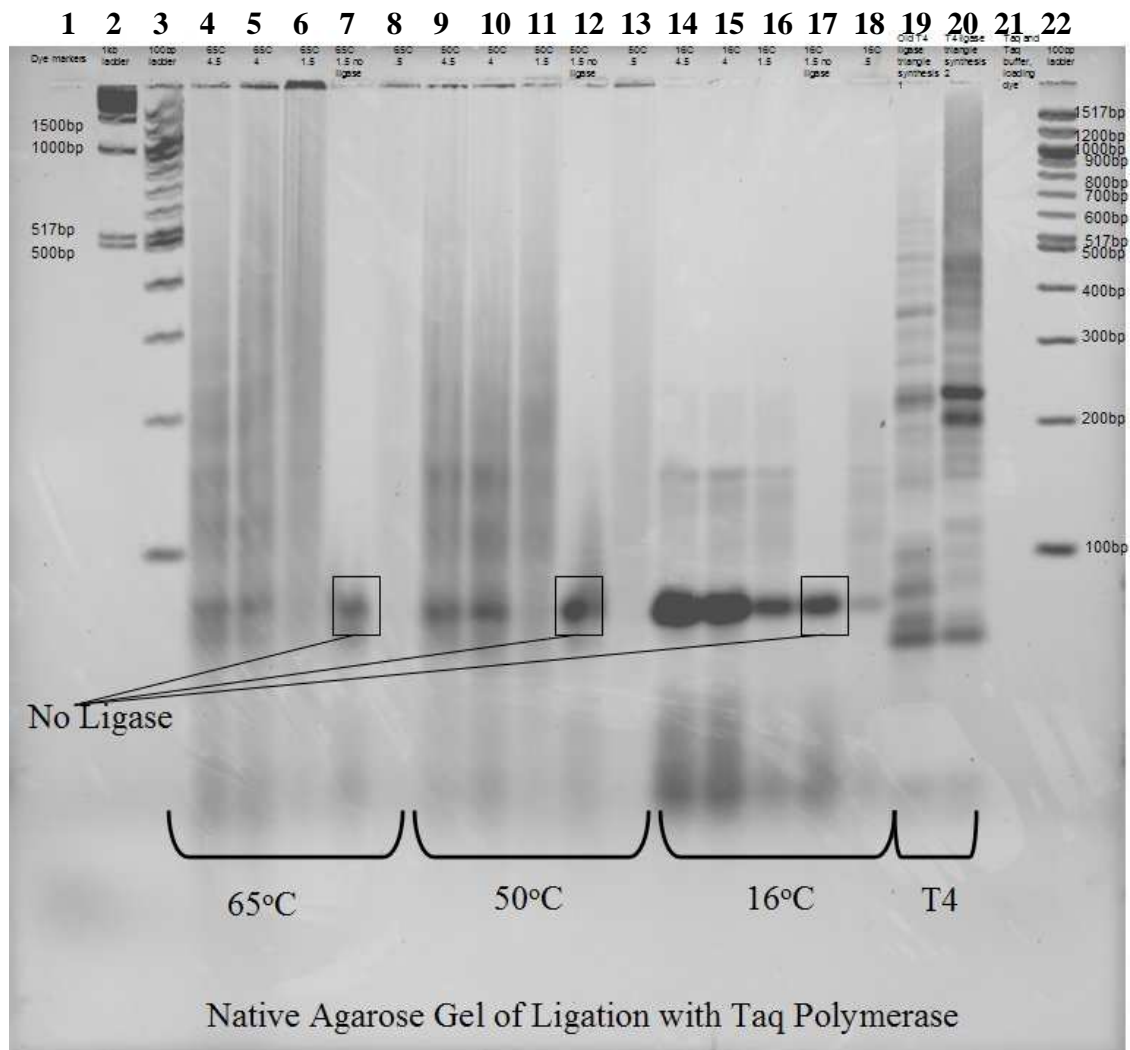
Figure 19: Native Agarose Gel of Taq ligation. Note the absence of triangular product bands, which should be near 117 bp. Lane one contains dye markers (xylene cyanol and bromophenol blue). Lane two contains a one kilobase DNA marker ladder. Lane three and twenty two contain a hundred base DNA marker ladder. Lanes four through six and eight show the ligation conducted 65° C, sampled at 4.5 hours, 4 hours, 1.5 hours, and 0.5 hours, respectively. Lane seven is a no ligase control at 65° C sampled at 1.5 hours. Lanes eight through eleven and thirteen show the ligation conducted 50° C, sampled at 4.5 hours, 4 hours, 1.5 hours, and 0.5 hours, respectively. Lane twelve is a no ligase control at 50° C sampled at 1.5 hours. Lanes fourteen through sixteen and eighteen show the ligation conducted 16° C, sampled at 4.5 hours, 4 hours, 1.5 hours, and 0.5 hours, respectively. Lane seventeen is a no ligase control at 16° C sampled at 1.5 hours. Lanes nineteen and twenty are two different ligation experiments with T4 ligase for comparison with Taq. Lane twenty one is a no DNA control, *i.e.* the Taq ligase and buffer without DNA.
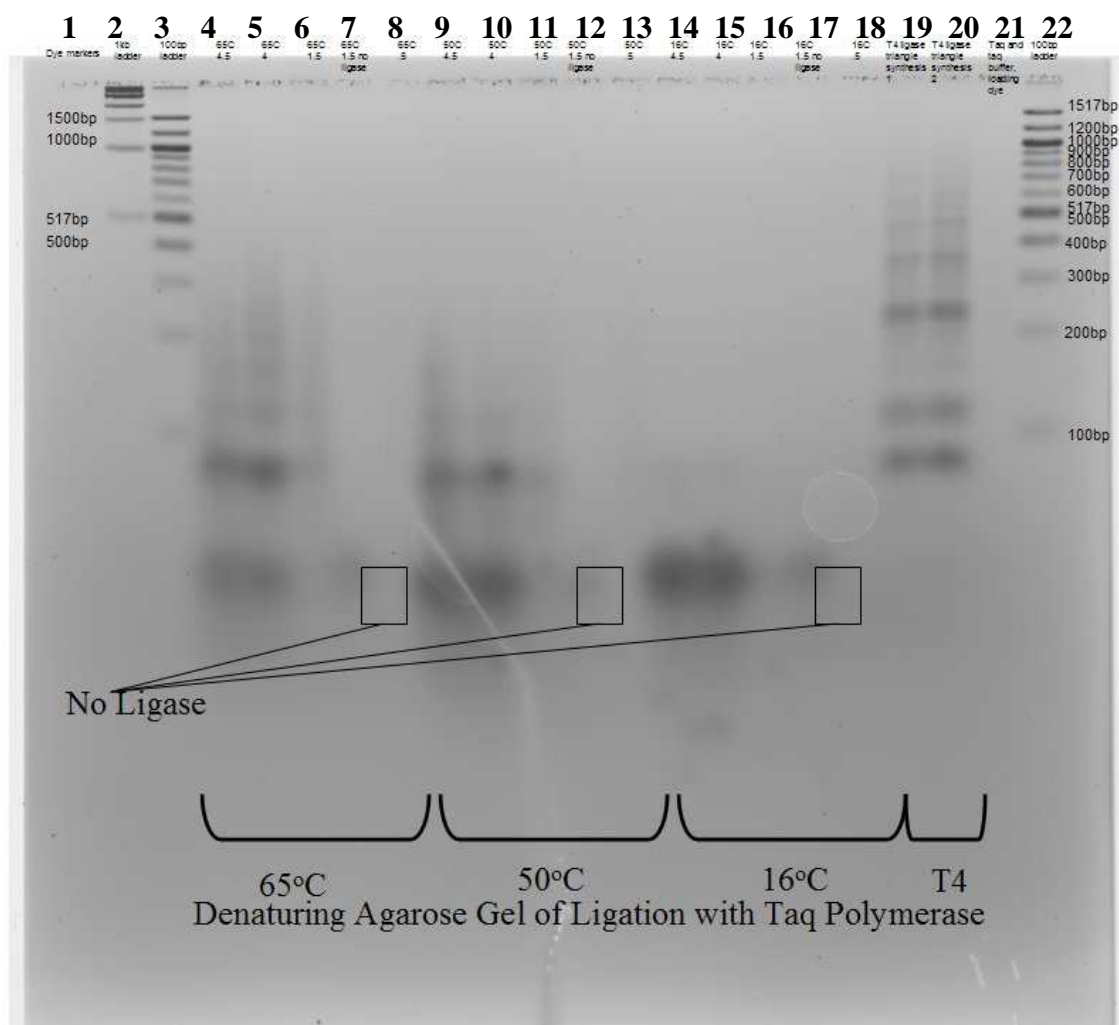
Figure 20: Denaturing Agarose Gel of Taq ligation. Note the absence of clean triangular product bands, which should be near 117 bp. Also note that some of the strands have been ligated (the bands greater in size than a 40mer). The lane identification is identical to figure 19. The denaturing is acomplished by alkaline conditions (50 mM NaOH, 1 mM EDTA)

Our interpretation of these gels is that while the desired association is taking place at the level of one side of the designed triangle (*i.e.* an 80 bp long piece), the whole structure is not forming effectively (this would be indicated by a 117 bp piece). The larger structures seen in the ligation involving the T4 ligase, as opposed to the Taq ligase, are due to the promiscuity of the ligase and the cool reaction temperature used ($16^{o}C$) which favor incorrect association. The problem may lie with the three way junctions that

we have engineered into the structure. If they do not have the assumed 60 degree angle, they will prevent the correct association instead of encouraging it. This is not a ligase dependent problem, as the strands do not associate correctly in the absence of ligase. One problem with the association could be a few problems with the base pairing design discovered after the assembly attempt. Some of our sequences have small areas of self complementary sequence, and portions which can base pair in a manner that can encourage T4 ligase to ligate the incorrect strands together. This was the reason for experimentation with Taq ligase. The Taq ligase experiments had no bands corresponding to fully ligated triangles. This eventually drove us to our experiments with chemical crosslinking. The problems in assembly shown by the gel electrophoresis data are reinforced by our AFM (Atomic Force Microscopy) data presented below in figures 21, 22, and 23.

AFM is a flexible technique that allows imaging on the nanoscale. At its simplest, AFM is a measurement of the height of a point on a surface. A very sharp point (the AFM tip) probes the surface and this point is scanned across the surface acquiring height information. The extremely small deflections of the AFM tip are measured by reflecting a laser beam off the reflective back of the tip. The reflected light is allowed to travel along a long path amplifying the angular deflection; much like a lever. Usually, a lever is used to transform a small force applied along a large distance, to a large force applied to a small distance. Using this principle in reverse, a lever can amplify motion. The amplified displacement of the beam is detected by a photodiode array. The position of the tip (x and y coordinates) and the height (the z coordinate) measured at that position are combined in a computer program to produce a topographical map of the area scanned by the AFM tip.

If the surface is very flat, and a nanostructure is attached to the surface, an image of the attached nanostructure can be obtained. For nanostructures as small as DNA, a very flat surface is required, so that the structure of interest is not lost in the background surface roughness. Luckily, a natural mineral, mica, has the interesting property of cleaving with atomically flat surfaces[31], and can be obtained cheaply in large enough pieces to be useful experimentally. The atomically flat surface of a cleaved piece of mica can then be used as a substrate for imaging DNA. As can be seen from the pictures, there are structures that, with imagination, are similar to our designed structure. However, they are the exception, and not the rule. Generally, amorphous structures are seen, indicating that our synthetic effort with ligase was inefficient.
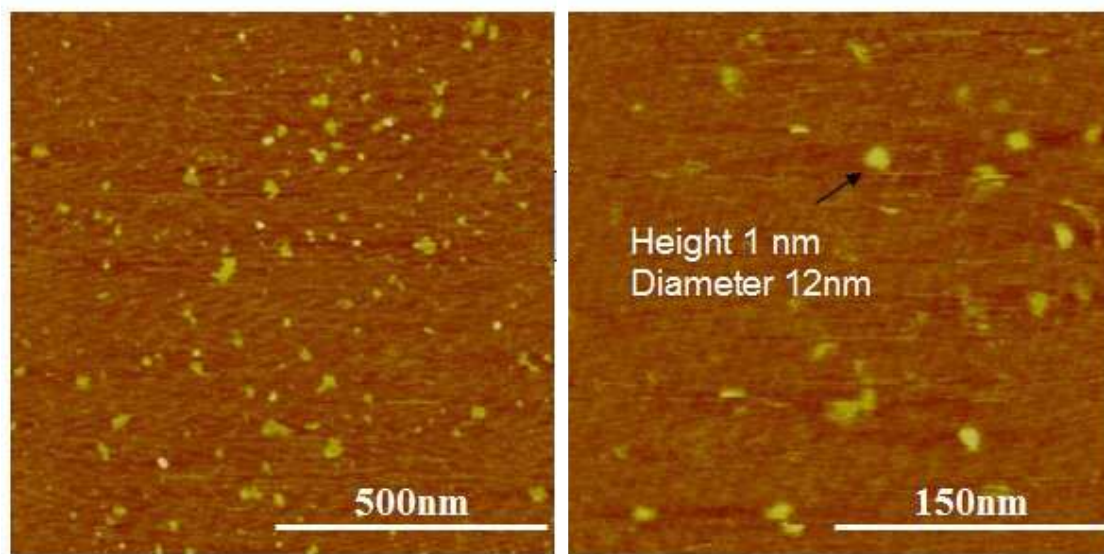
Figure 21: AFM Image of the Original Triangle Synthesis with T4, Unpurified by Gel Electrophoresis. Note the wide variety in sizes of the structures imaged, indicating heterogeneity of the sample.
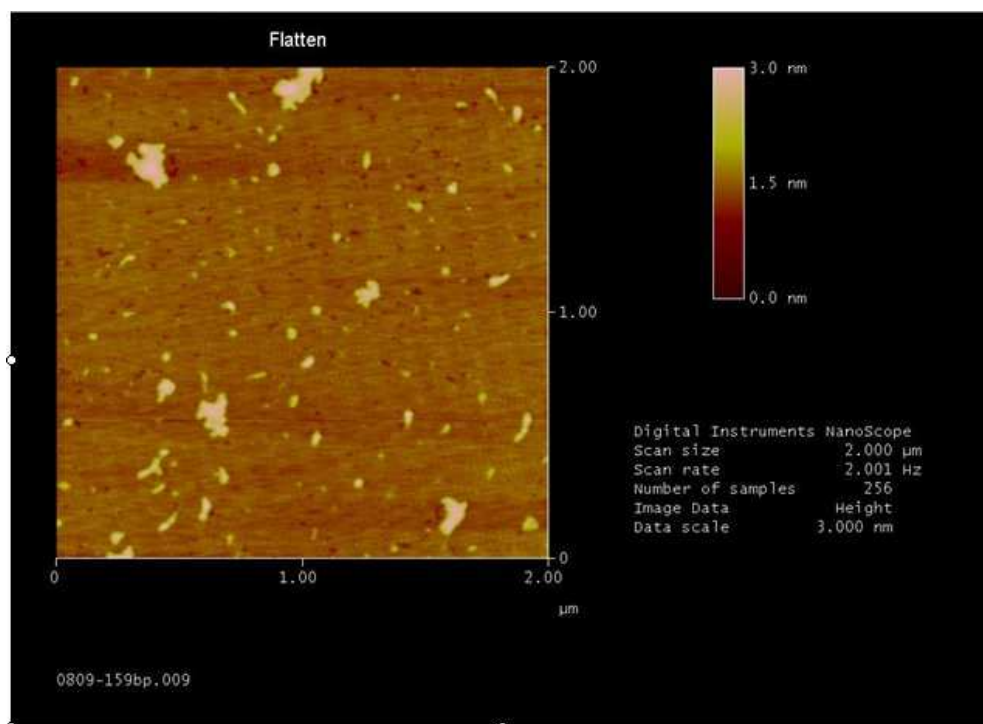
Figure 22: Image of the Purified 117mer Band, T4 Ligation. Note the wide variety in sizes of the structures imaged.
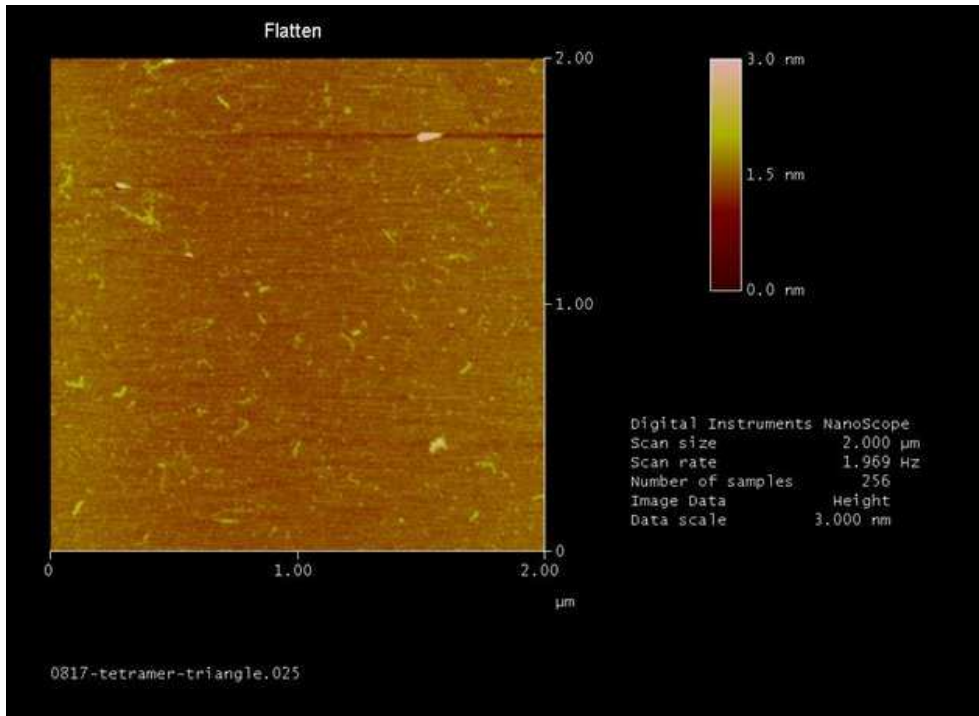
Figure 23: Image of the Purified 468mer Band, T4 Ligation. Note the wide variety in sizes of the structures imaged.

3.3.2 Crosslinking

The problems found with our design for ligase assembly, drove us to look for other ways to construct the triangle. Our attention turned to different methods of locking our strands together. Chemical crosslinking appeared a good choice. If the crosslinking was controllable and reversible it could be used much like "tack" welding in assembling steel structures. It would act as temporary glue to hold pieces together, to allow piecewise assembly, and when the structure was complete they could be taken out. This would also allow purification in intermediate steps (not our original goal, but would allow optimization, where our previous results, *i.e.* no positive result, would not). Our first choice was the crosslinker mechlorethamine. Mechlorethamine can crosslink DNA at the

N7 position in guanine, preferentially in 5' GNC regions[32]. It is a relatively small crosslinking agent and introduces minimal distortion into B-form DNA[32-33]. 5' GNC regions in our design were found in all strands, save one. In our hands, (see figure 24) mechlorethamine proved to be an inefficient crosslinker, is not easily reversible, and safety and handling proved onerous (Mechlorethamine is a vesicant, derived from sulfur mustard chemical warfare agents, and due to its cytotoxic properties is sometimes used for   chemotherapy).
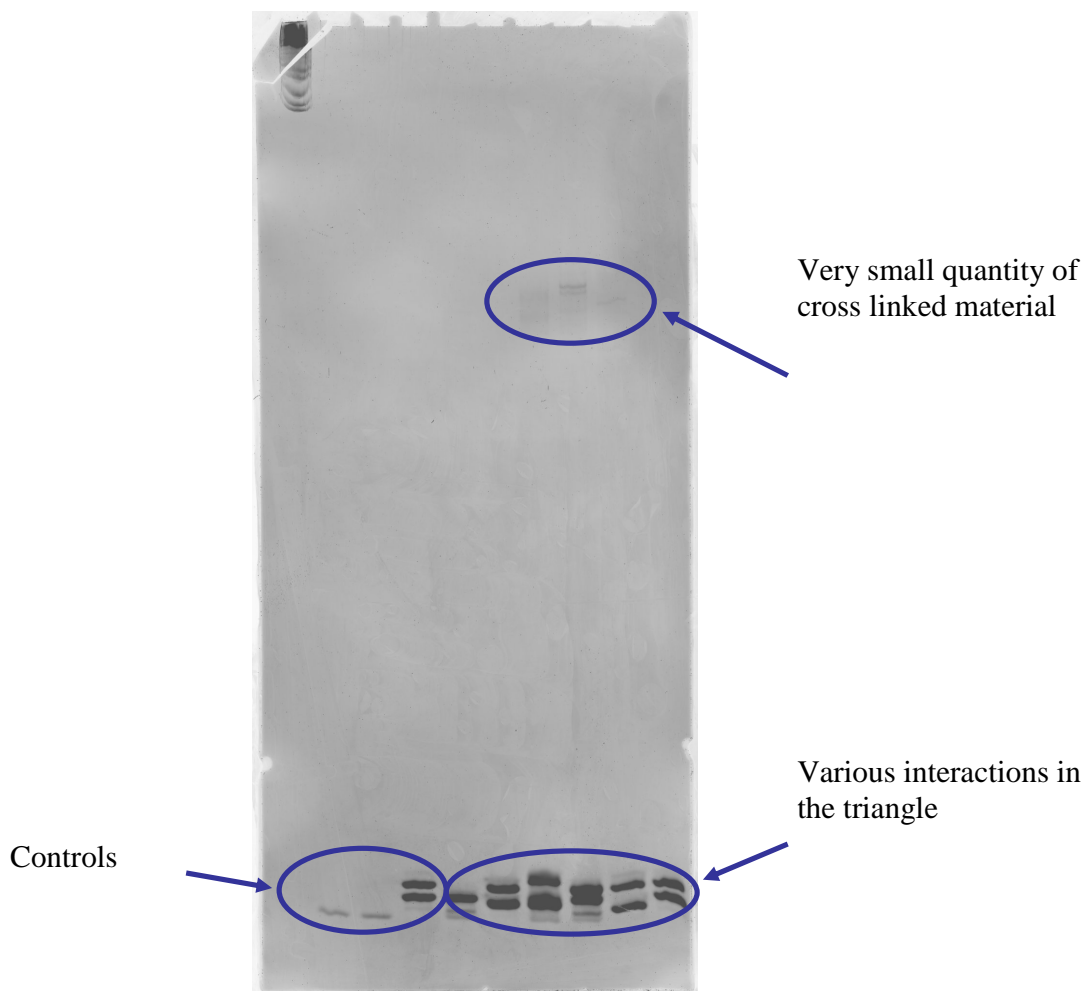
Figure 24: Denaturing polyacrylamide gel of mechlorethamine addition to annealed triangle components. Note the lack of full sized triangle.

Mechlorethamine was then abandoned in favor of psoralen.
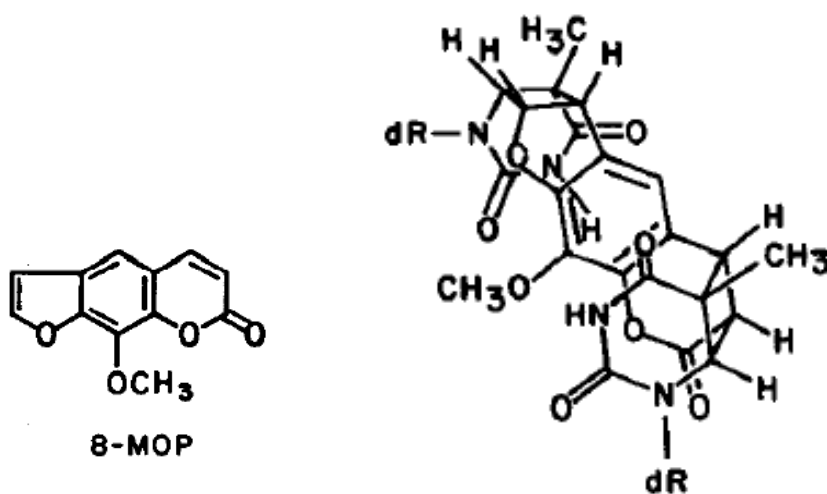
Figure 25: The structure of 8-methoxypsoralen[34], and the cyclobutane linkages created between the thymines in the preferred AT steps.
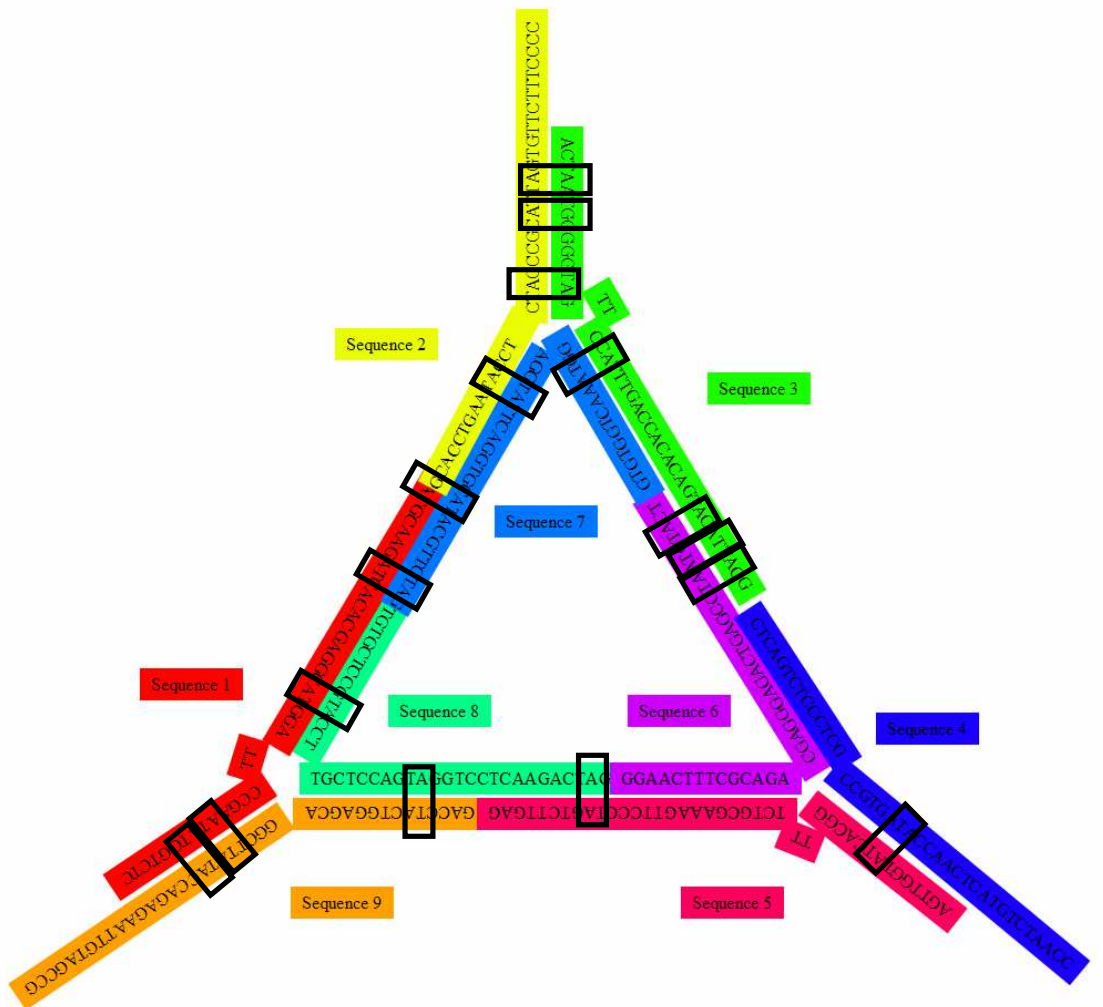
Figure 26: The junctions in the DNA triangle that are cross-linkable with psoralen

Psoralen, a chemical derived from diseased celery[35], seemed to be a superior choice as a crosslinker. 8-methoxypsoralen (one of the many psoralen derivatives available),or 8-MOP, was chosen as it was a photosensitive crosslinker, reversible, is still used in clinical practice to treat intractable skin conditions, is available cheaply, and had sequence specificity for AT repeats. It is, conveniently, safer to handle than mechlorethamine. The psoralen intercalates between base pairs in the double helix. There is then a stronger driving force for association of crosslinker and nucleic acid, unlike the

case of mechlorethamine, which is probably depends on weak, nonspecific electrostatic interactions to bring it to crosslinkable positions[36]. We had reason to expect that this association, before the crosslinking step, would improve crosslinking efficiency. Also arguing for its selection was that AT repeats exist in all of the overlaps between the strands that connect the triangle together, save one (see figure 26). After association, 8-MOP forms crosslinks between the two strands of double helical DNA *via* pyrimidine bases under the illumination of UV light in the 340 nm range[37]. UV light in the 240-313 nm range[38] will break the crosslink.

We decided to start with a piecewise "tack-welding" assembly of our structure with psoralen (see figure 27).
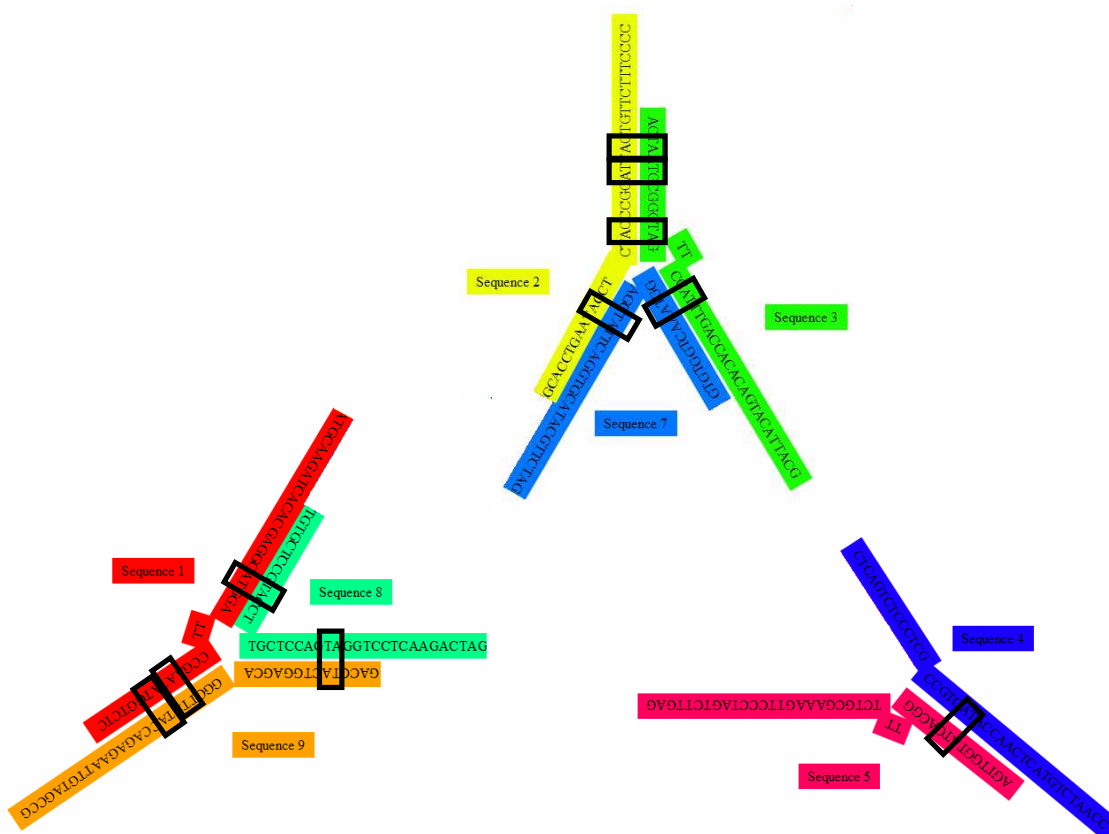
Figure 27: The vertices that were cross linked in the first step of an assembly plan with 8-MOP

 This was done because our one pot synthesis with ligase or mechlorethamine had many unwanted products, and did not assemble into the desired structure. The stepwise synthesis and purification would allow dissection and diagnosis of the problems, and complete the assembly.

We started with the three way junctions or vertices in the planned triangle. On denaturing polyacrylamide gels, we could identify the desired product bands, as well as unexpected undesired products. Sequence seven in the triangle had a small area of self complementary base pairing, and importantly, the AT step preferred by 8-MOP. The thermodynamics for association for this small amount of base paring are very low in energy, but do happen transiently. This transient association seemed to be enough to crosslink the 2-3-7 junction to itself through strand seven forming a dimer of the vertex. We theorize that 8-MOP intercalation can stabilize the AT steps, before crosslinking making the transient association longer lived, allowing more time for crosslinking.
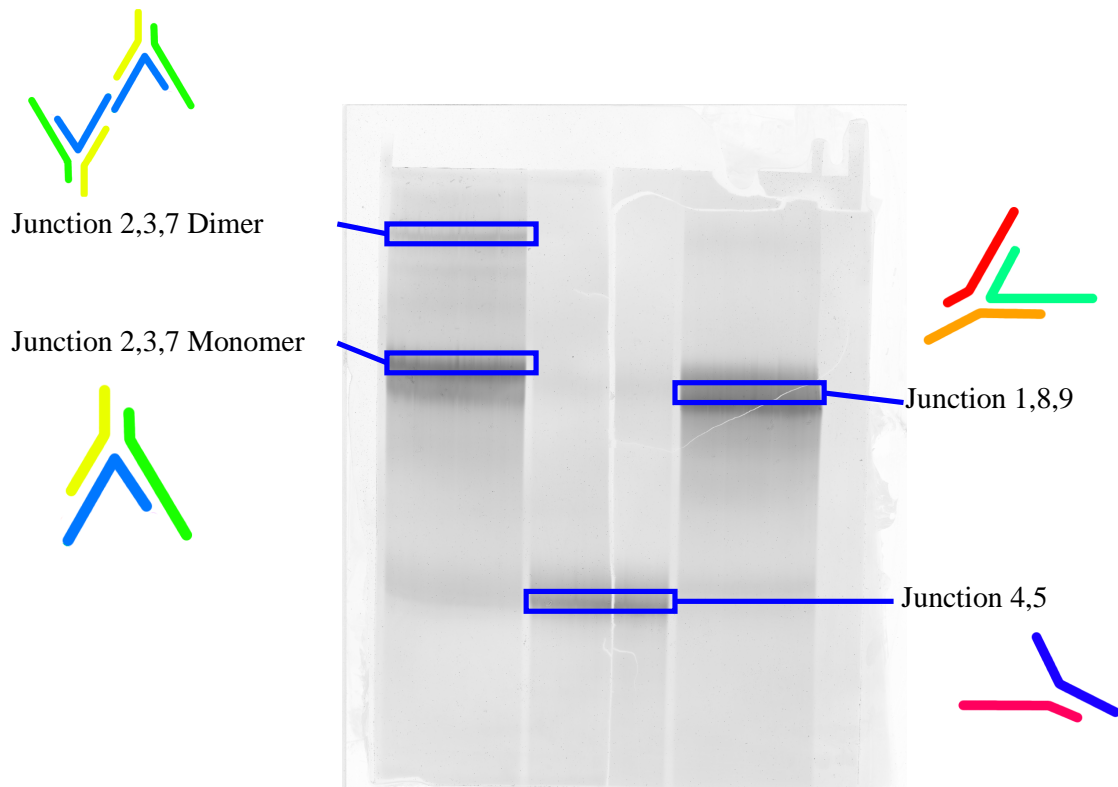
Figure 28: Denaturing PAGE gel of the vertices in the triangle. The unassociated strands have been run off the bottom of the gel. The proposed interactions for each band are labeled.

The size of the vertices was verified with matrix assisted laser desorption ionization mass spectrometry (MALDI-MS). MALDI-MS employs a solid matrix that can be partially ionized, and converted to a gas by a UV laser. The molecule of interest is embedded in the matrix, and with the matrix support changed in phase from solid to gas, is swept into the gas phase by the matrix plume, and then the mass spectrometer. Charge transfer from the matrix to analyte allows separation in the mass spectrometer. This is a relatively gentle ionization technique. Mass spectrometry of large nucleic acids however is a challenge. Charged in the gas phase, they tend to depurineate, fragment, and form phosphate salts, creating wide mass distributions and reducing the reliability of mass determinations[39]. Reliably determining the mass of nucleic acids longer than 100 base

pairs has not been successful to date. Our fully assembled triangle is larger than this limit. This explains our poor mass spectra displayed in figure 34. We did obtain reliable masses for the junctions and suggestive data for posited larger assemblies.
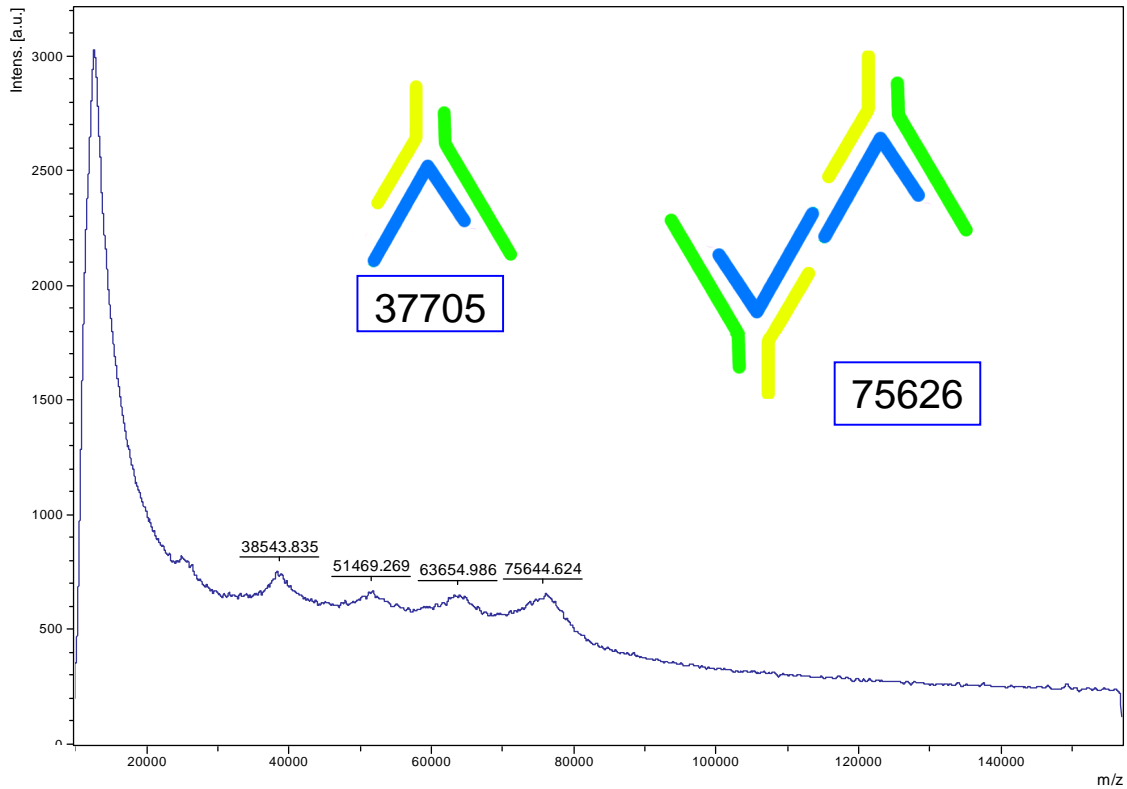
Figure 29: MALDI-MS of strands 2, 3 and 7, crosslinked with 8-MOP, the posited dimer band excised from denaturing PAGE gel. . The calculated mass for the assembled parts are boxed, and the measured masses are underlined.
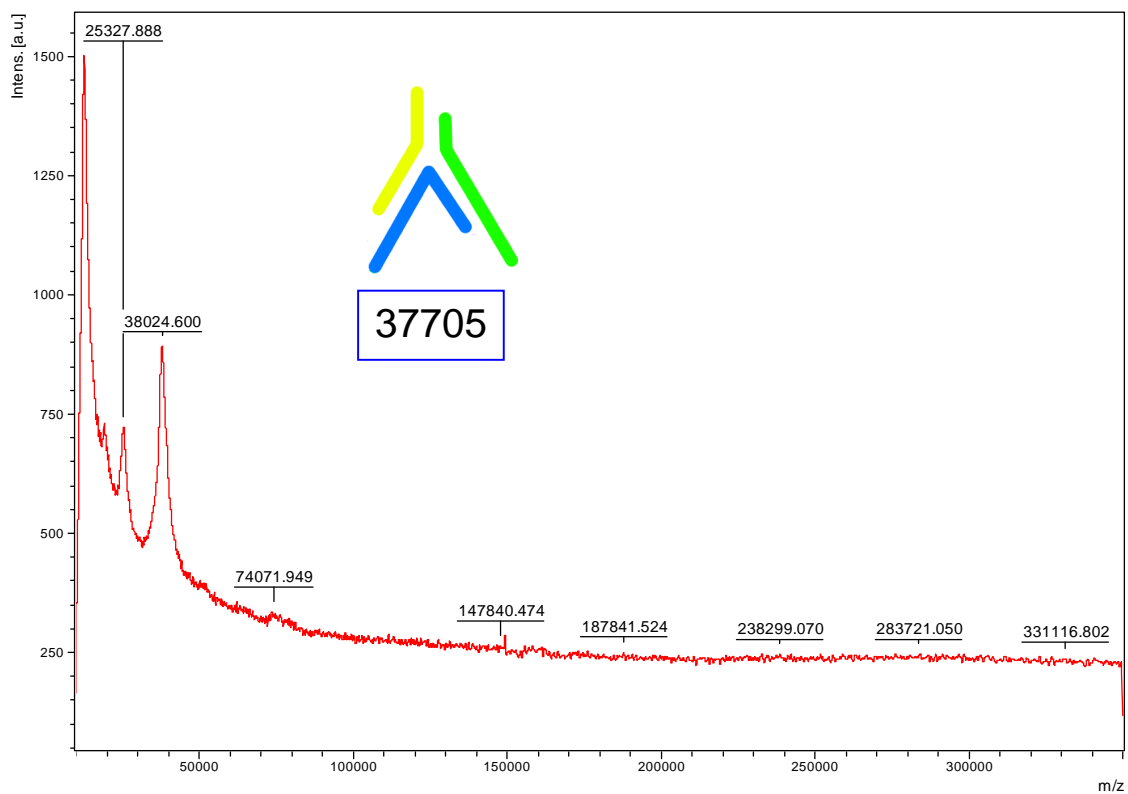
Figure 30: MALDI-MS of strands 2, 3 and 7, crosslinked with 8-MOP, the posited monomer band excised from denaturing PAGE gel. . The calculated mass for the assembled parts are boxed, and the measured masses are underlined.
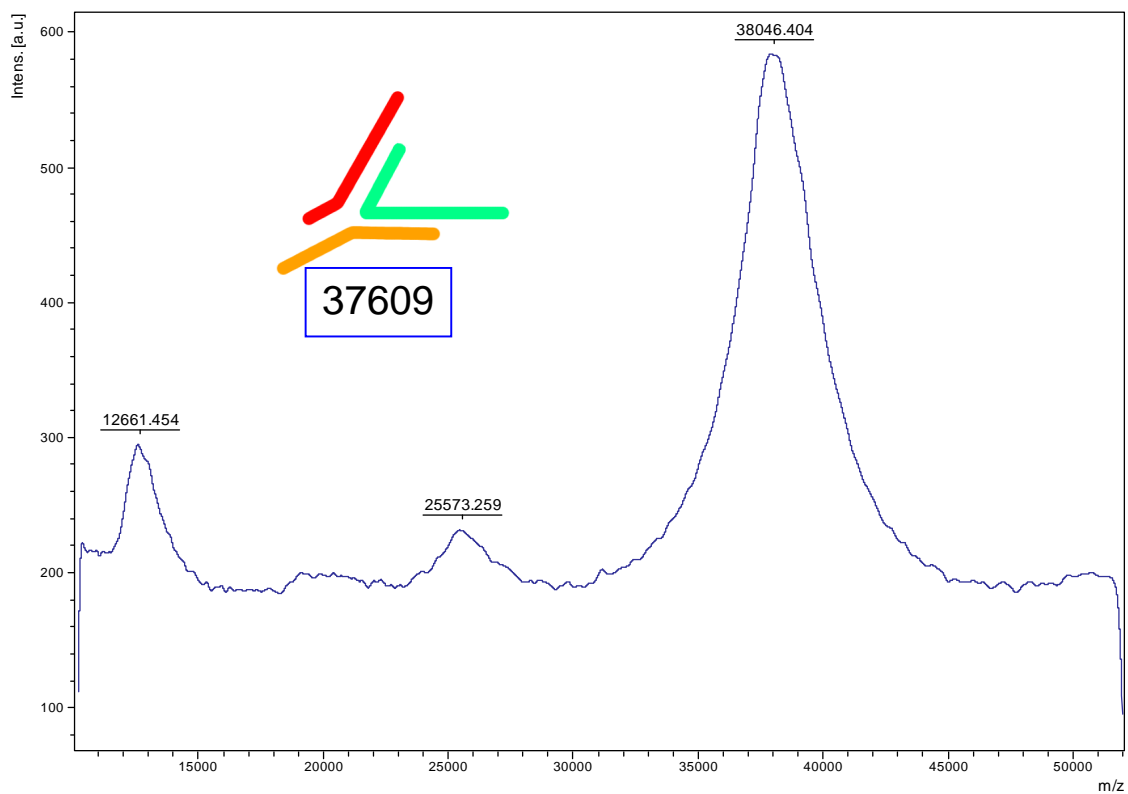
Figure 31: MALDI-MS of strands 1, 8 and 9, crosslinked with 8-MOP. . The calculated mass for the assembled parts are boxed, and the measured masses are underlined.
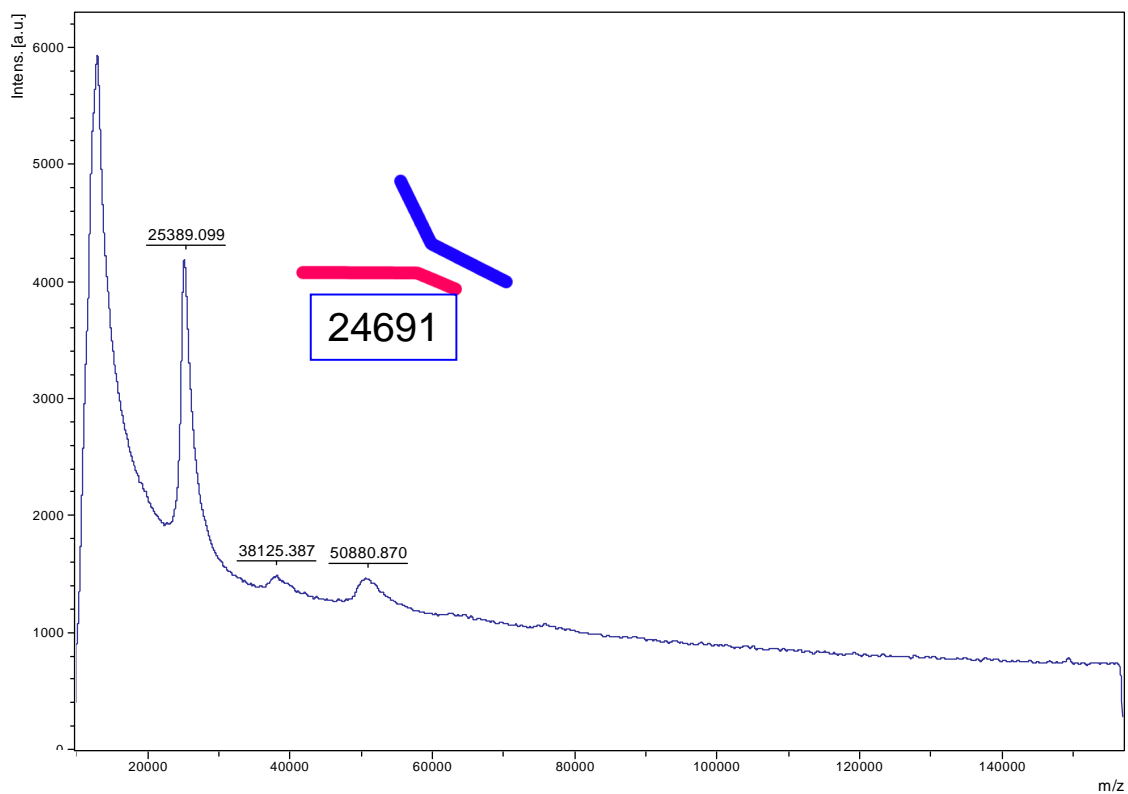
Figure 32: MALDI-MS of strands 4 and 5, crosslinked with 8-MOP. . The calculated mass for the assembled parts are boxed, and the measured masses are underlined.
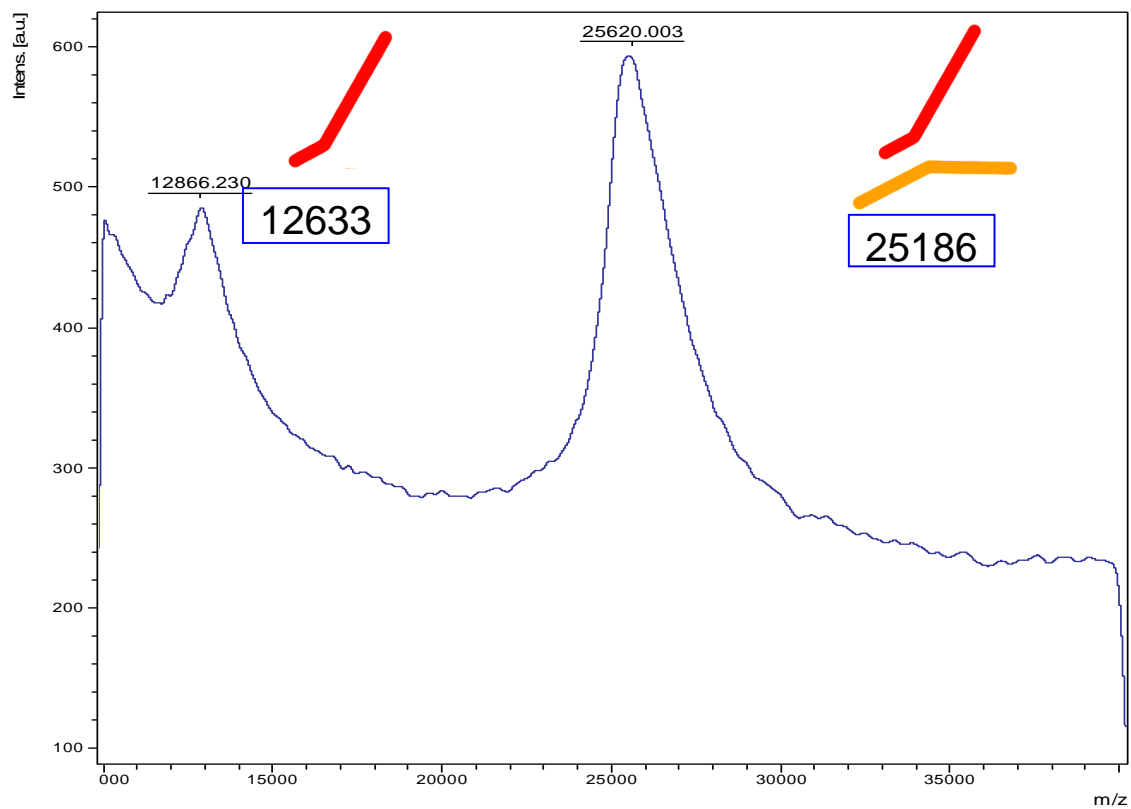
Figure 33: MALDI-MS of strands 1 and 9, crosslinked with 8-MOP. . The calculated mass for the assembled parts are boxed, and the measured masses are underlined.
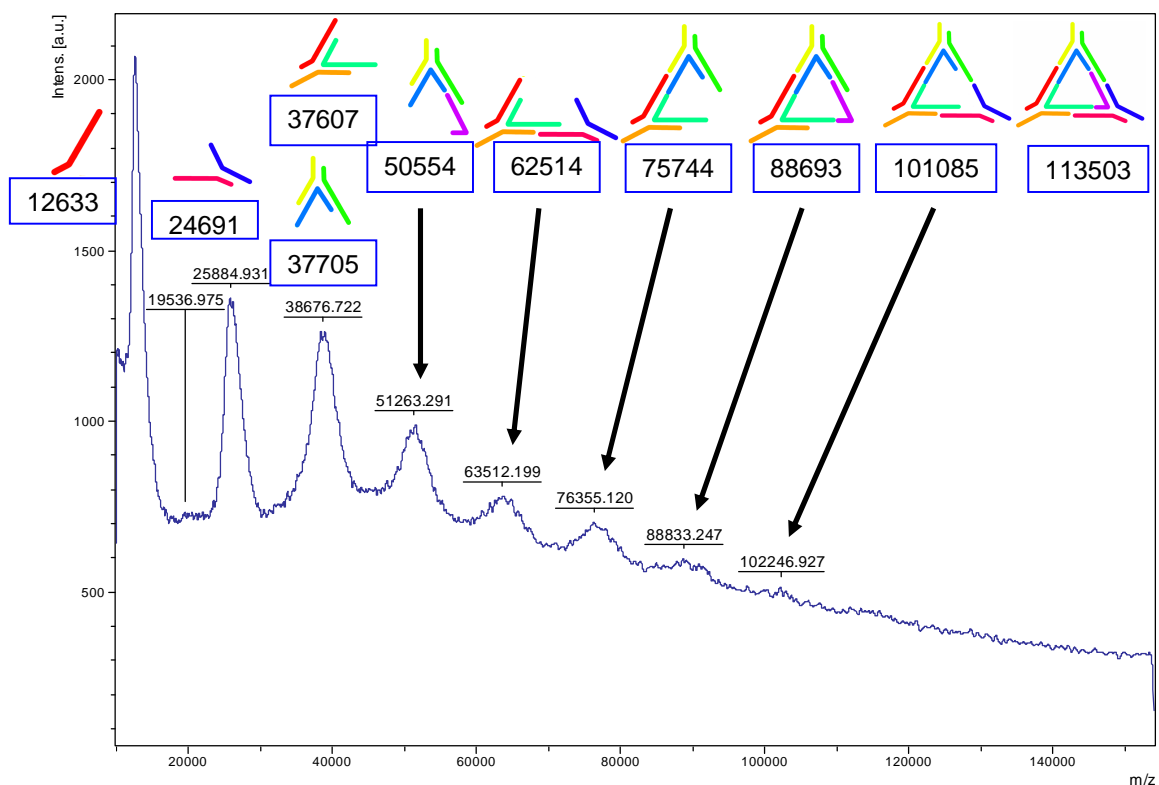
Figure 34: MALDI Mass spectra of assembly of the triangle from junctions. Masses corresponding to single strands rather than the junctions are most likely due to the crosslink breaking in the ionization process. The calculated mass for the assembled parts are boxed, and the measured masses are underlined. The analyzed sample, in all likelihood, is made up of junction sized pieces, and larger assemblies; the measured intensities correspond more directly to stability of pieces of that mass to charge ratio.

The next step was the assembly and crosslinking of the proposed full size triangle. The denaturing PAGE gel indicates larger products that are consistent in size with the designed triangle. Interestingly, a dimer product is also shown. This is probably through self-complementary portion of strand 7, shown to associate and crosslink by our vertex experiments. As the mass spectra for the larger assemblies is inconclusive we turned to another technique, dynamic light scattering. DLS relies on the fact that particles of any size scatter light. If the intensity of scattered light over time is monitored over time, for a

small illuminated volume, the intensity of scattered light is seen to fluctuate. The fluctuation is due to particles drifting in and out of the measurement volume. The Stokes-Einstein equation can describe how spherical particles diffuse dependent on their size. If this equation is used to fit the autocorrelation function of the fluctuations in intensity over time, the size of diffusing particles can be deduced. The general idea is that smaller particles diffuse faster, so the intensity of scattered light from the measurement volume is autocorrelated for a smaller period of time than that of larger, more slowly diffusing particles. This means that the measured "size" is dependent on the hydrodynamic radius. This procedure works well for spherical or ellipsoidal particles, but the theory for other geometries is still being developed. For our triangular designs, there is some indication that we would see two distributions, corresponding to the edge or face of the triangular structure.
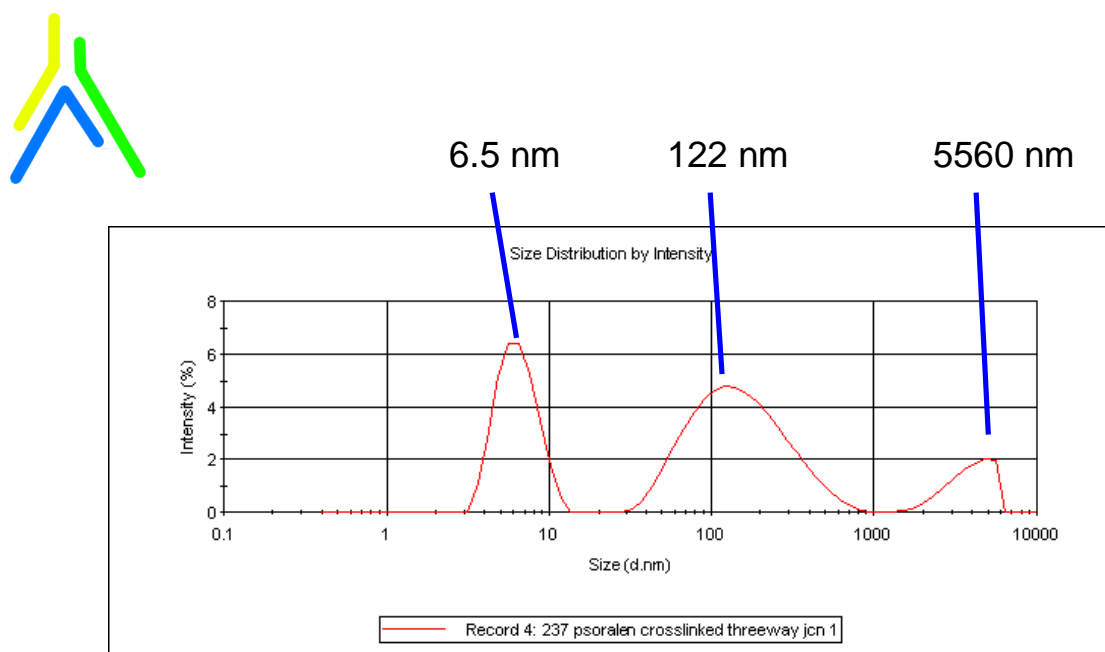
Figure 35: DLS intensity distribution for the 2, 3, 7 junction
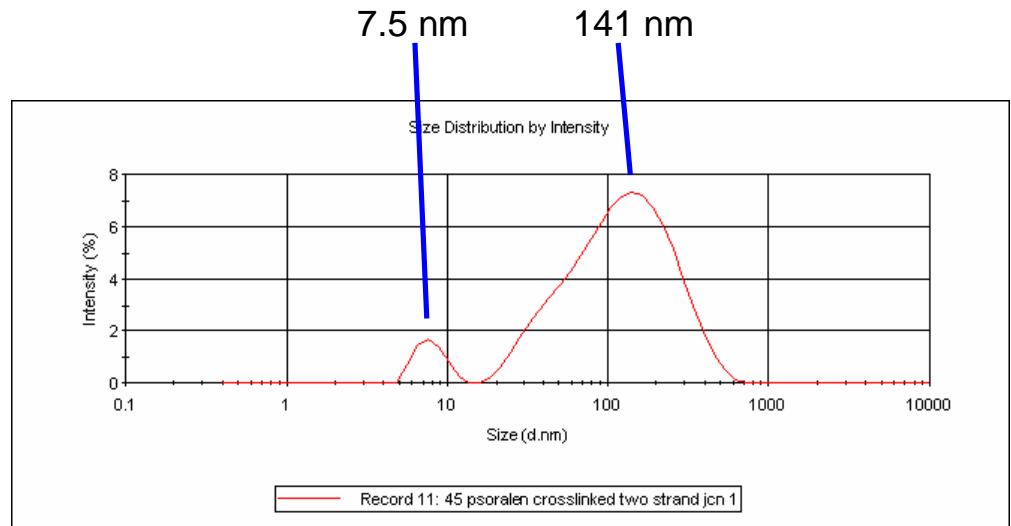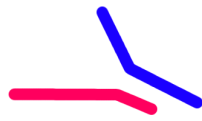
Figure 36: DLS intensity distribution for the 4, 5 junction
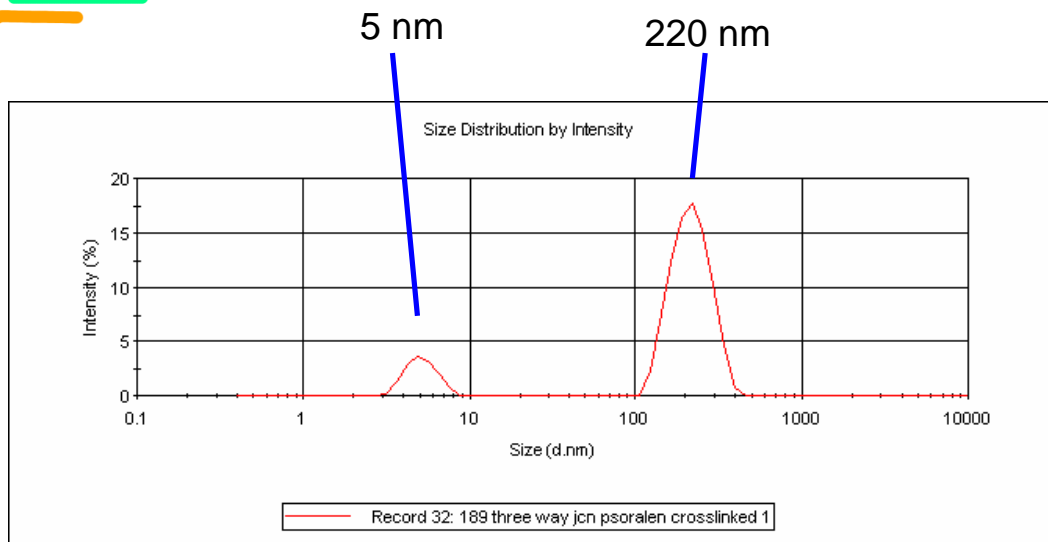
Figure 37: DLS intensity distribution for the 1, 8, 9 junction
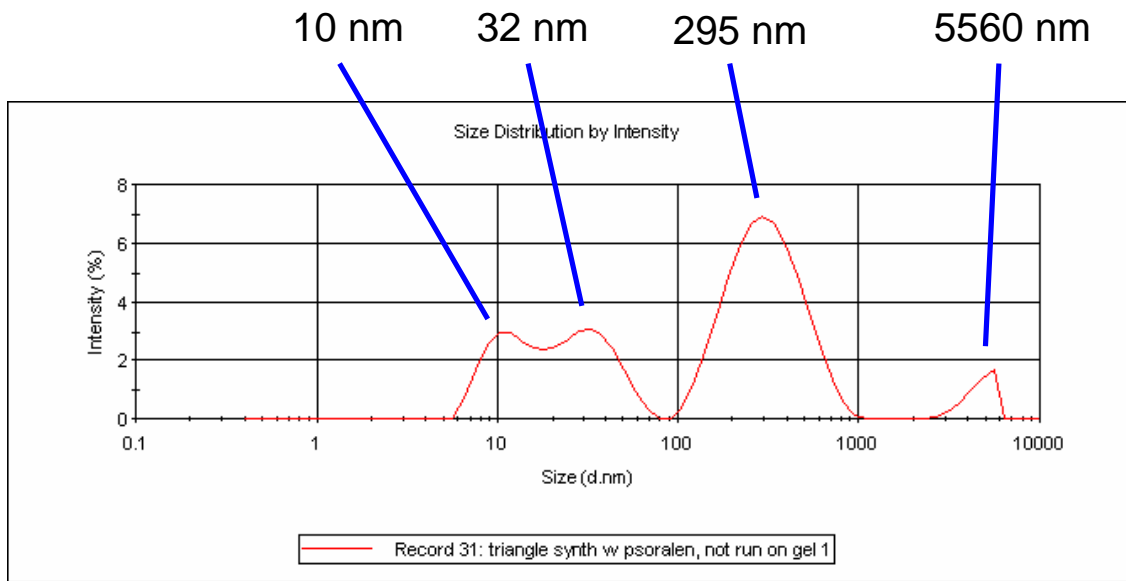
Figure 38: DLS intensity distribution for synthesis of the whole triangle, from junctions
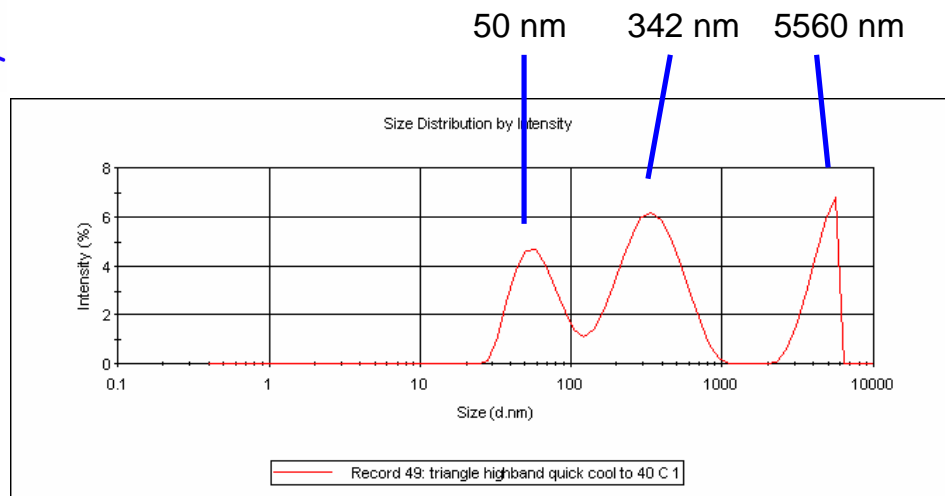
Figure 39: DLS intensity distribution for the proposed dimer triangle band from denaturing PAGE of the complete synthesis.
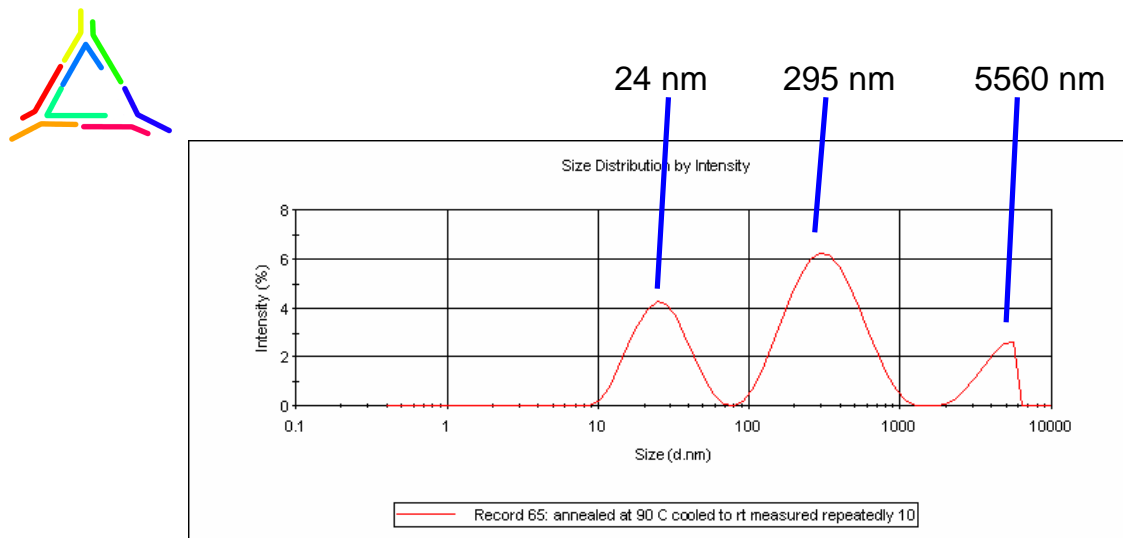
Figure 40: DLS intensity distribution for the proposed monomer triangle band, missing one strand, from denaturing PAGE of the complete synthesis.
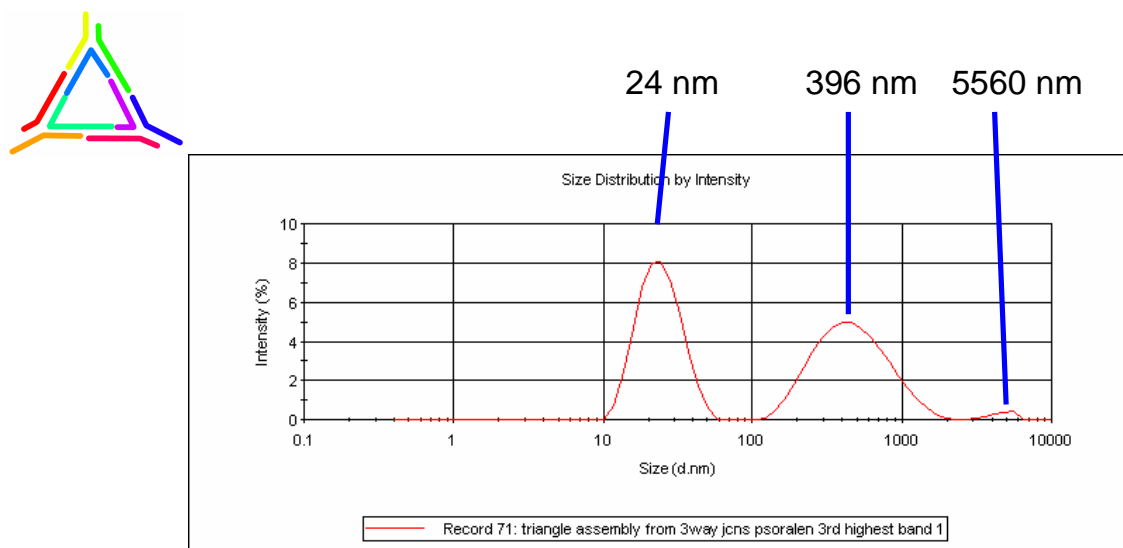
Figure 41: DLS intensity distribution for the proposed monomer triangle band from denaturing PAGE of the complete synthesis.

As can be seen from the measured sizes from DLS, they corresponded well to the predicted dimensions for the triangle. The larger measured size may correspond to the scattering for the face of the triangle as opposed to the edge. The largest species probably corresponds to bubbles forming in the solution during observation. With this encouragement, we again employed AFM for structural detail.

y: 0
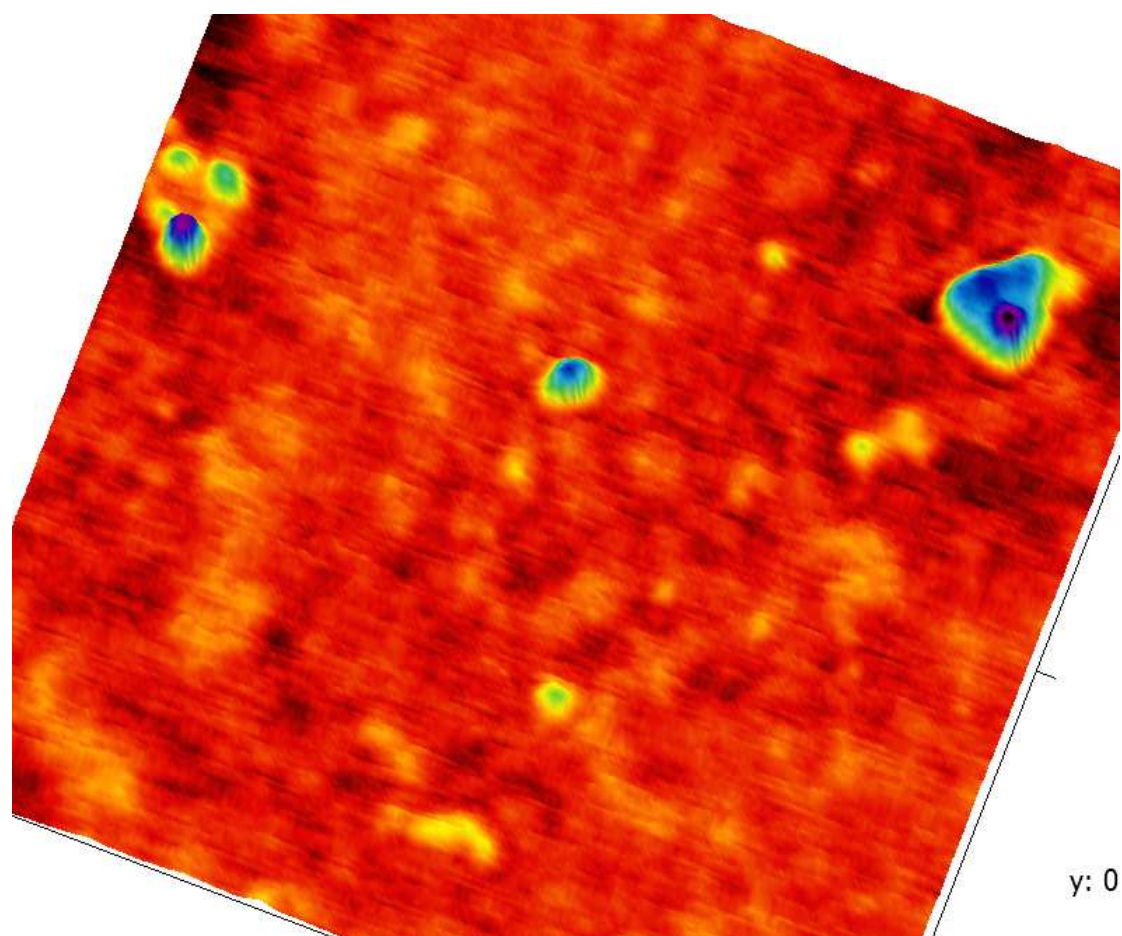
Figure 42: AFM image of psoralen cross linked, gel purified sample corresponding to the designed triangle. The dimensions and geometry correspond to the model.
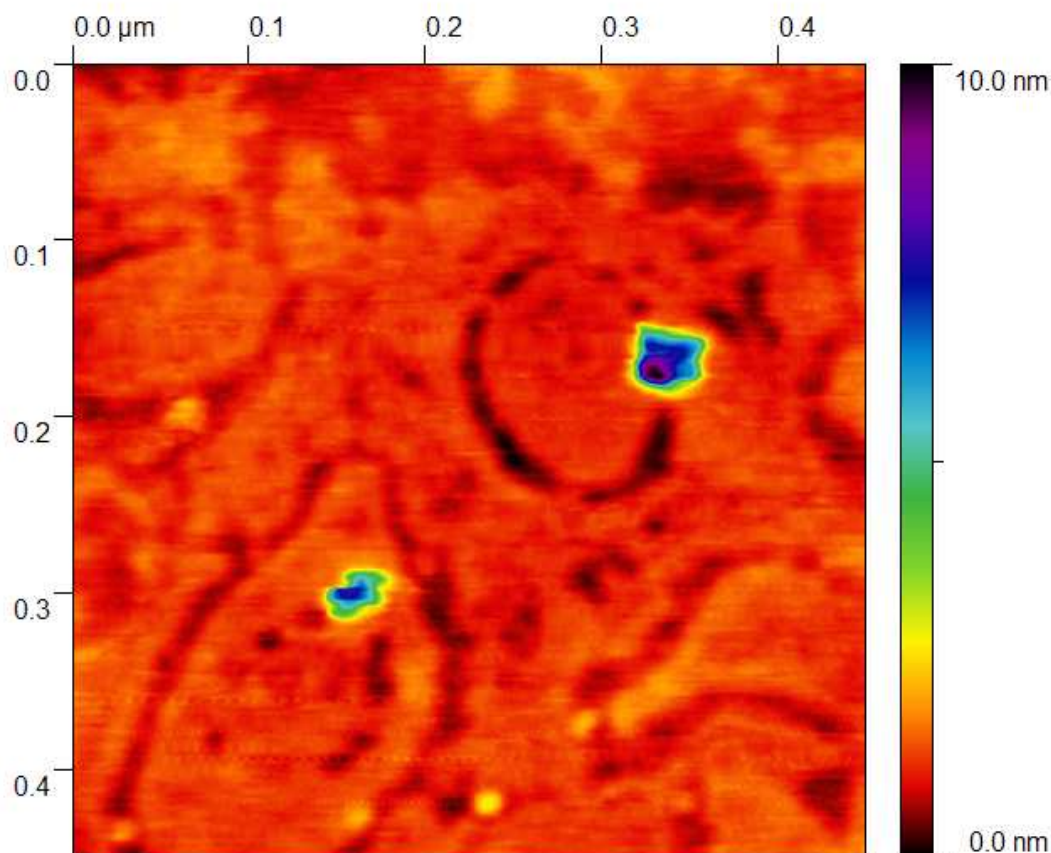
Figure 43: Another example AFM image of psoralen cross linked, gel purified sample corresponding to the designed triangle. The dimensions generally correspond to the model, but the geometry does not.

3.4 Conclusion

As seen in figures 42 and 43, much like our work with the ligase, we only found a few (3-4 examples, depending on what one considers a triangle) examples of structures that agreed with our design. While our gel and light scattering data are consistent with the designed structure, we do not see enough of the desired structure with AFM to draw any conclusions, except for perhaps failure in design. When one looks long enough at the molecular level, one can see structures that look like almost anything. To fairly determine if what is observed is different from these imaging artifacts; many hundreds of examples

in close proximity are needed. This very small population of structures that correspond to the design may be due to the one connection that is not crosslinkable, and because of this the structure was unstable in the relatively harsh conditions employed in our AFM imaging (The RNA is desalted, and dried on a mica surface, this can subject a sample to large forces[40]). The larger structures would be then due to oligomerization through this open connection. Another possible explanation is that AFM imaging on a mica surface is an extractive technique[41]. We may have inadvertently employed conditions that selected aggregates, and incorrectly assembled structures, or single stranded structures[42]. This may not have been detected in MALDI due selection of smaller species and the possibility of the MALDI ionization process breaking up aggregates. Another problem may be that small quantities of incorrectly assembled structures may act as catalysts to nucleate large scale aggregation and misassembly (most probably through the non-crosslinkable junction). This nucleation would make sense if other structural concerns prevent correct association. The phase of the helix or the angle of the junctions, while designed to be compatible for association (and to prevent oligomerization), may not be correct for complete triangle formation. These issues warrant further exploration by other researchers, as our thermodynamic calculations are known to be reliable, and if our methodology could be made to work would allow much more freedom in DNA nanostructural design. In the end, the question of success of our design methodology is unanswered. We found that we had to give up our original motivation of one pot synthesis, efficiency and good yields to make any progress with the assembly. The gel and light scattering data are encouraging, but the AFM data is not. Perhaps a new design,

informed by the problems encountered with this first design could meet with more success or the use of different imaging strategies (x-ray crystallography, SAXS, *etc.*).

4. RNA Regulatory Dynamics

4.1 Introduction

Setting aside our work in accurate structural design and synthesis, dynamics are also necessary to construct functional nanomachines. To gain some understanding of how nucleic acids interact dynamically with other cellular components we chose the TRAP protein. The TRAP protein is a classic example of a protein nucleic acid interaction, and can undergo many different types of interactions. We hoped to use this as inspiration for dynamic machinery based on nucleic acids.

4.2 The TRAP/Tryptophan/RNA Regulatory System

The tryptophan regulatory attenuation protein or TRAP regulates tryptophan synthesis in many *Bacilli*[43]. The TRAP protein is involved in regulation of the tryptophan biosynthetic genes at least four different points [44]. Three of these regulatory points involve only the protein, mRNA and tryptophan [44a, b, 45] at the transcriptional and translational level. The last involves another protein (the anti-TRAP protein) and couples the system to not only tryptophan levels, but the levels of charged tRNA[trp44e, 46]. The TRAP protein regulates the expression of the *trp* operon transcriptionally by binding to the leader element of the operon in a tryptophan dependent manner. The binding of the protein causes a restructuring of a large secondary structure formed by the leader element. After restructuring, a normally occluded terminator hairpin forms and signals the polymerase to stop transcription. Alternatively the first gene in the *trp* operon, *trpE,* contains another structure that, upon TRAP binding, hides the Shine-Delgarno sequence, preventing transcription[44a]. The TRAP protein can also directly compete with ribosome binding in the case of the *trpP*[45] (a postulated tryptophan transporter)and *trpG*[44b] (a

glutamine amidotransferase common to both the folic acid and tryptophan pathways) genes (also *ycbK,*[47] postulated to be an efflux protein) The fourth way in which the TRAP protein is coupled to the tryptophan system in the cell, is if charged tRNA[trp] levels drop sufficiently, the anti-TRAP protein is expressed. This protein can bind to TRAP, preventing it from binding RNA, regardless of tryptophan levels[44e, 46a-c].
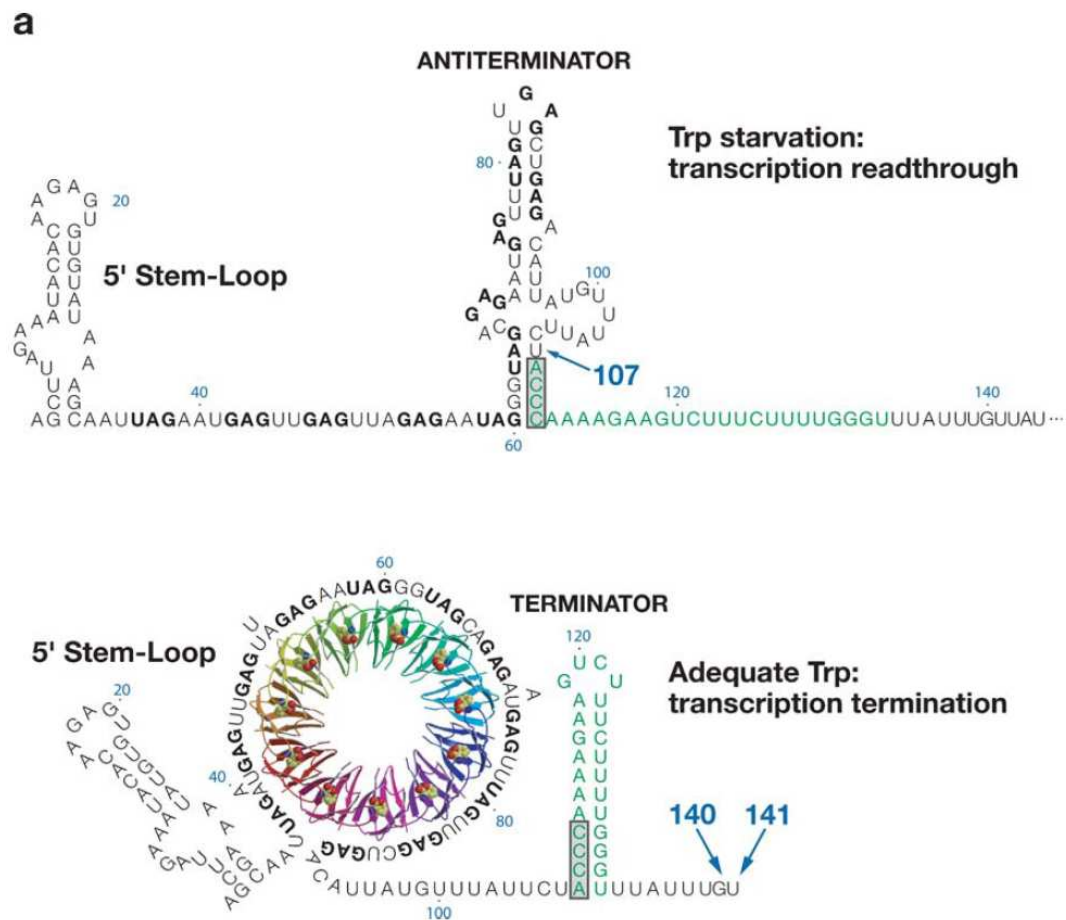
Figure 44:Transcriptional regulation by the TRAP protein[48]. In the absence of tryptophan, the leader RNA forms a RNA structure called the antiterminator. This structure does not interfere with the progression of transcription and the tryptophan bio-synthetic proteins encoded in the *trp* operon are completely transcribed. In the presence of high tryptophan concentrations, the TRAP protein binds to the leader element rapidly, before the polymerase progresses to the protein encoding regions of the operon. The binding causes a restructuring of the antiterminator, allowing a standard terminator hairpin to form. The terminator signals to the polymerase to stop translation and disassociate from the DNA, never having transcribed the protein encoding regions of the operon. Thus, tryptophan levels in the cell decrease, due to lower levels of tryptophan synthesizing proteins.
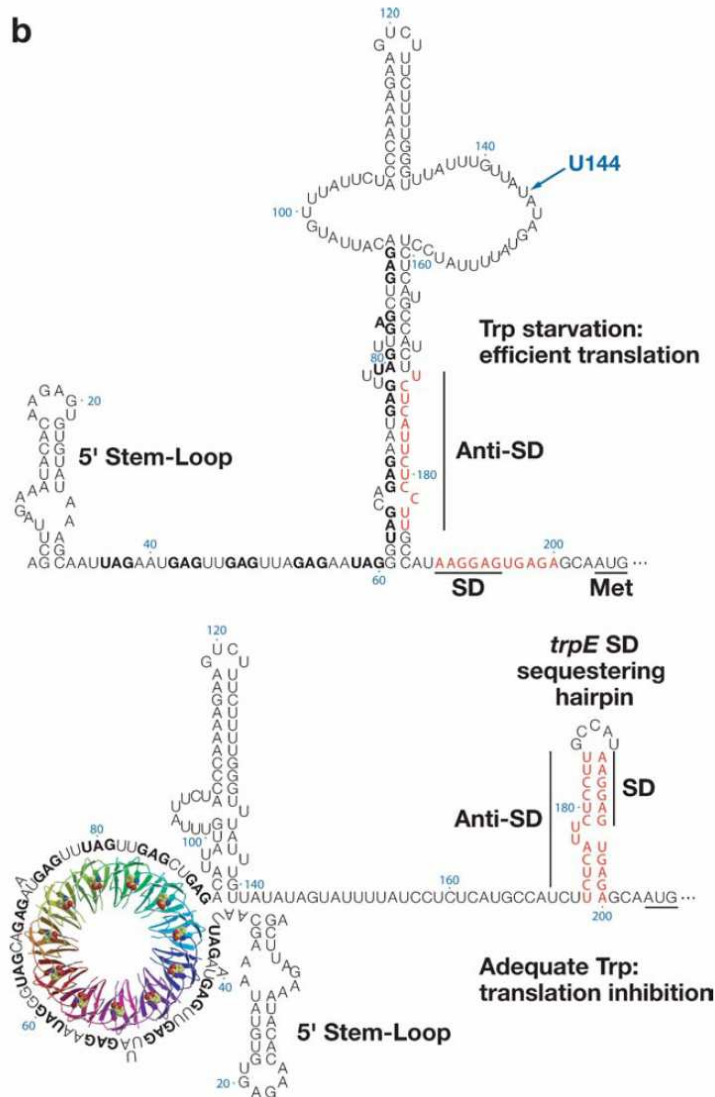
Figure 45: Translational regulation by the TRAP protein[48]. The TRAP protein has a second chance to interfere with the synthesis of protein if the complete mRNA is transcribed. If tryptophan levels are low, another larger RNA structure is formed in the leader area of the mRNA. This structure has an exposed Shine-Delgarno sequence; the ribosome can bind to the Shine-Delgarno sequence and initiate translation normally. In the presence of high tryptophan levels, the TRAP protein can again bind and reorganize this structure. In this case, the reorganization sequesters the Shine-Delgarno sequence in a hairpin structure. The ribosome can no longer bind and translation of the operon does not occur. Again, lower levels of tryptophan synthesizing proteins, lead to lower cellular levels of tryptophan. This secondary regulatory function may be because tryptophan levels can change between transcription and translation, or perhaps because the translational regulation is imperfect.
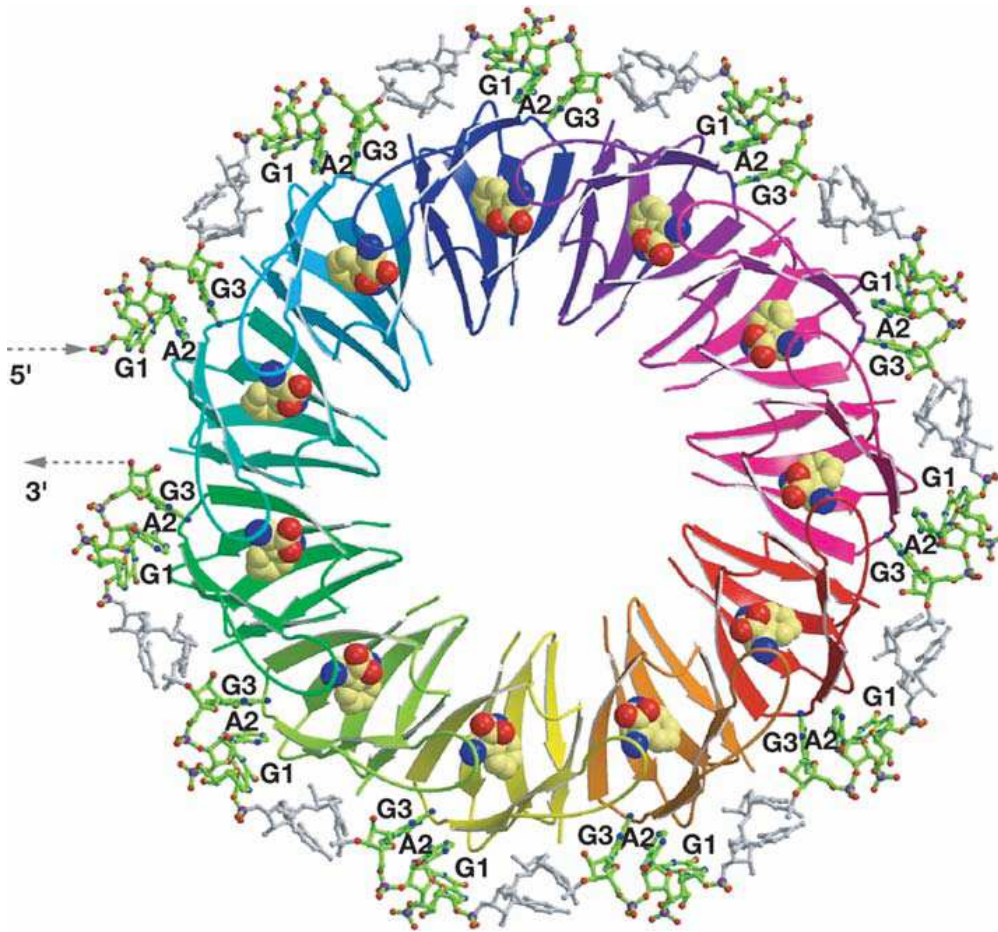
Figure 46: The crystal structure of the TRAP protein (represented by the rainbow coloured ribbons), fully saturated with tryptophan(represented by the CPK coloured spheres), with RNA bound(represented by the CPK coloured ball and stick model).[48-49]

The TRAP protein, itself, is composed of eleven identical subunits, which assume a toroidal or ring shaped structure, of total molecular weight of approximately 91,000[49-50]. Each subunit in the TRAP protein binds to UAG or GAG repeats separated by two nucleotide spacers[51] if activated by tryptophan binding. An interesting property of the system is that tryptophan binding to the protein is non-cooperative[52] to weakly cooperative[53], depending on the species of origin (TRAP from *B. subtilis* seems to bind cooperatively, but TRAP from *B. stearothermophilus* seems to be non-cooperative) but protein binding to the RNA seems very cooperative[49]. Also, there is a preference for 4-6

units of the RNA binding motif[54]. After approximately this number of repeats the protein shows little increase in affinity[54]. The TRAP protein also seems to not undergo large scale concerted conformational changes[55]. Cooperative systems offer organisms a method of generating tunable and switch like effects in regulation and binding and so are advantageous. However the details of how a weakly cooperative to non-cooperative process in tryptophan binding to the TRAP protein results in a very cooperative process in protein binding to RNA have not been explained. We have performed single molecule FRET experiments to probe this interaction. We believe that our data suggests an explanation for this effect. The data, qualitative pictures of the interaction, and a mathematical model are presented. Our models elaborate on an evolutionary pathway for multimeric allosteric proteins. By following our suggested pathway backward with retro-mutations, we suggest that an effort in "mutational archeology" may be an interesting avenue of research. Finally, this evolutionary path suggests a strategy for nanomachine design.

4.3 Experimental Design

As bulk experiments have not sufficiently explained the behaviour of the TRAP system, a single molecule approach was taken. Single molecule experiments can provide detailed information on dynamics, show transient intermediates, and give information on the temporal ordering of required states[56]. This information is often not available or obscured in bulk measurements. We chose FRET to report on the state of the single molecules. Florescence or Förster resonance energy transfer is a powerful spectroscopic technique used to measure distances in biomolecules. This distance measurement relies on the non-resonant energy transfer from one fluorescent molecule to another. This

transfer can happen if the emission spectrum of one molecule overlaps with the absorption spectrum of another molecule. This is a sign that the energy levels in the respective molecules are matched, i.e. the energy difference between the ground state and the excited state in the donor molecule after vibrational and rotational relaxation is equal to the gap in the acceptor molecule. The efficiency of energy transfer is strongly dependent on distance, and so can be used as a "molecular ruler". The energy transfer efficiency depends on distance in a one over the distance to the sixth power manner, giving good sensitivity between 2.5nm to 10 nm. To use FRET in our system, we installed two fluorophores that were FRET paired into a synthetic mimic of the leader element of the *trp* mRNA (the site of TRAP protein binding for transcriptional regulation). The fluorescent labels were on either end of the leader RNA so that they could monitor the length of the RNA during binding by the TRAP protein. In the unbound state there would be a low FRET intensity due to the large distance between the fluorophores. In the bound state, there would be a high FRET intensity as the TRAP protein brings the ends of the RNA (and the fluorophores) together. Finally, the RNA construct was also biotin labeled, to allow surface immobilization and single molecule observation.

Figure 47: The RNA construct used in the experiments. FRET paired fluorophores (Cy3 as the donor and Cy5 as the acceptor) are attached to each end of a synthetic RNA construct designed to mimic the TRAP binding site. We expect a low FRET efficiency when the construct is unbound and the ends of the RNA are far apart. When the TRAP protein binds, it will bring the ends of construct together, resulting in a high FRET efficiency (see figure 46 for the crystal structure of the TRAP protein bound to RNA, and figure 48 for diagrams of the different states).Biotin is attached to the 3' end (the right side of the construct) to allow for surface immobilization. Surface immobilization allows long observation of individual molecules (see figure 48)
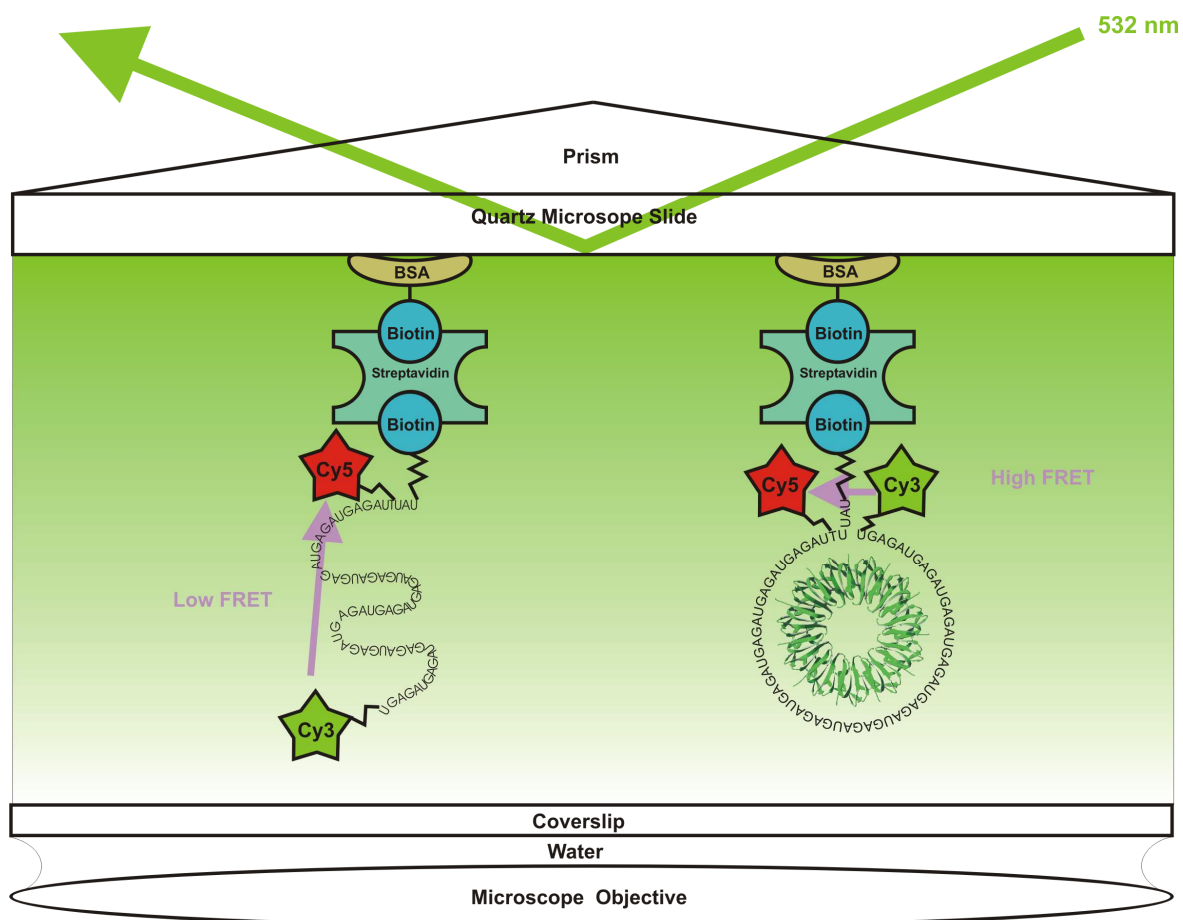
Figure 48: Experimental design, surface immobilization, and expected FRET values for our system. The RNA construct is immobilized on the quartz surface of a microscope slide through the biotin-streptavidin interaction. First, bovine serum albumin (BSA) with covalently attached biotin is introduced to the slide. The BSA non-specifically absorbs to the surface of the slide. Streptavidin is then introduced and binds to the biotin. Then our construct is added to the slide. As streptavidin has four binding sites for biotin, our biotin labeled RNA can bind to the streptavidin, and through this be tethered to the surface. The left cartoon depicts the immobilized RNA without bound TRAP protein, giving a weak FRET signal (low energy transfer efficiency, therefore low acceptor emission intensity). The right cartoon depicts the immobilized RNA with bound TRAP protein, giving a strong FRET signal (high energy transfer efficiency, therefore high acceptor emission intensity). The surface immobilization serves two important purposes. Firstly, it allows observation of the molecules over long periods of time. Secondly, it allows excitation of a very small portion of the sample compartment by TIRF (total internal reflectance fluorescence). In fluorescence experiments, background fluorescence is usually a problem, more so in single molecule experiments where the desired signal is weak. The excitation laser is totally internally reflected off the top of the slide. This generates an evanescent wave in the solution at the interface. This wave rapidly decays, (the intensity is negligible a few hundred nanometers from the interface) but can excite donor molecules that are near the interface, like our surface immobilized construct. In this way most of the solution is never illuminated by the excitation laser, strongly reducing background fluorescence.

Generally, experiments were conducted and data was collected as outlined in figures 48 and 49. Our construct was immobilized to a quartz microscope slide by a three step process. First, biotin labeled bovine serum albumin (BSA) was non-specifically absorbed to the surface of the slide. Second, the immobilized BSA was bound by streptavidin. As streptavidin has four binding sites available, there are free sites for more biotin to bind. In the last step our biotin labeled RNA construct was bound to the streptavidin, linking it to the slide surface through the streptavidin and BSA. After this immobilization step, the RNA construct was exposed to the various conditions in the experiment (figures 50-55). The slide was illuminated with a 532 nm laser (this wavelength is absorbed by Cy3, but not by Cy5), and the resulting fluorescence from the two fluorophores was recorded by a CCD camera (see figure 48). The motion picture recorded by the camera was processed with custom computer code in the IDL programming language. These IDL programs extracted out the fluorescence intensity data for individual molecules from the movie, and as both the Cy3 and Cy5 wavelengths were recorded simultaneously, but spatially separated, and also matched these together. After this extraction process, the individual single molecule traces were analyzed for relevance with custom computer code written in the Matlab programming language. This included, making sure that the purported single molecule trace was only a single molecule (by one step photobleaching), and that artifacts were not incorrectly identified as molecules. Once this was completed the single molecule traces were summed into a histogram to display the overall behaviour of the observed molecules. The summing process was done with in house software written for the IGOR pro[57] computer program.
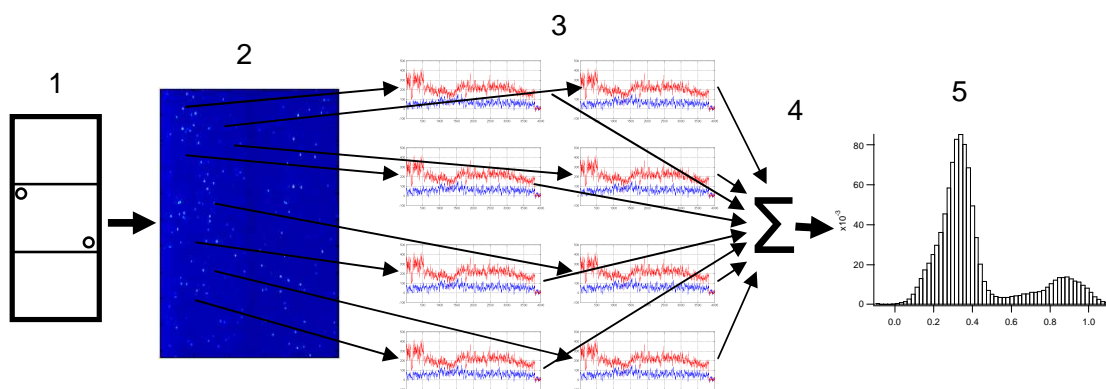
Figure 49: General single molecule workflow. First, (1) fluorescence data is collected from immobilized molecules on slide by a microscope and camera. Second, (2) from the movie file recorded by the camera, individual single molecule fluorescence intensity traces are extracted and the two different wavelengths recorded (for each molecule) are matched together custom IDL computer code. Third (3) the extracted traces are analyzed for relevance to the experiment with custom Matlab computer code. Fourth (4) the traces are summed together to finally create a histogram (5) of the overall behaviour of the single molecules. This summing is done with custom IGOR pro computer code.

4.4 Results

4.4.1 RNA alone condenses with increasing divalent concentration

Our first experiment was to assess the response of the RNA construct alone. Along with determining the behaviour of RNA without protein, divalent cations are known to cause tertiary structural formation[58]. Our construct was designed to mimic the natural leader element of the *trp* operon and while there is no published information showing structural formation in this area alone, with the large role RNA restructuring plays in the TRAP regulatory system, it was necessary to test this possibility. These two goals were accomplished by observing the behaviour of the RNA alone with various Mg2+ concentrations (from zero to sixteen mM).
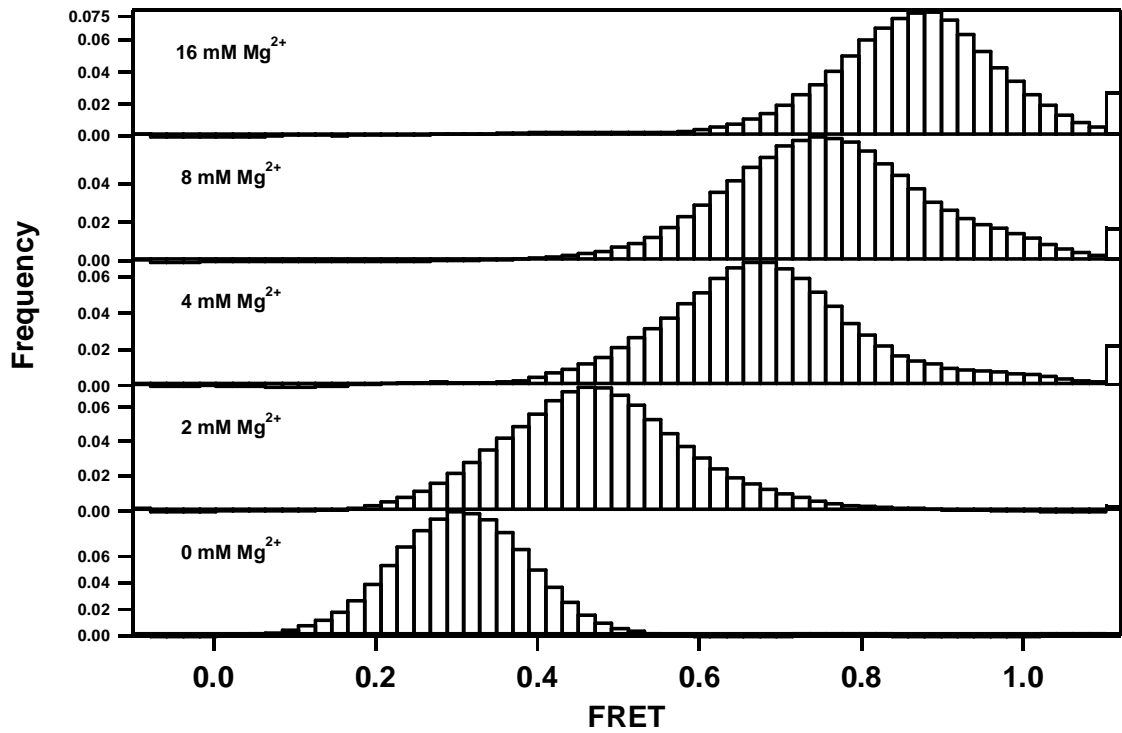
Figure 50: Histogram showing the behaviour of the RNA construct in the absence of the TRAP protein and magnesium. These experiments were conducted in 50 mM MOPS, pH 7.5, 100 mM NaCl, PCD/PCA[59] as the oxygen scavenging system, and 2 mM Trolox[60] as a antioxidant and triplet state quencher.
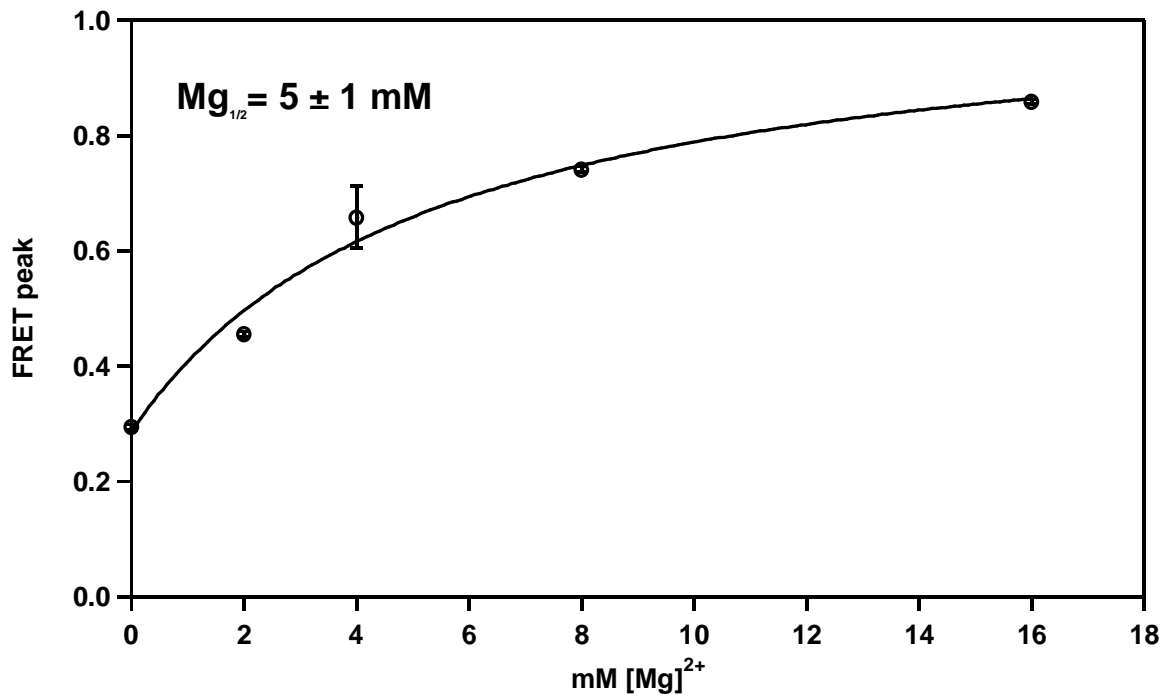
Figure 51: Binding curve for the RNA construct in the absence of the TRAP protein and magnesium. These experiments were conducted in 50 mM MOPS, pH 7.5, 100 mM NaCl, PCD/PCA[59] as the oxygen scavenging system, and 2 mM Trolox[60] as a antioxidant and triplet state quencher.

As seen in figures 50 and 51 the RNA construct in the absence of the TRAP protein responds to increasing Mg^{2+} concentration. This response takes the form of a gradual increase in the apparent FRET state dependent on magnesium concentration. We did not detect discrete individual states or shifting between different states in the single molecule traces used to make up the histograms. We suggest that magnesium is stabilizing normally unstable secondary or tertiary structural interactions in a nonspecific manner. These could take the form of reducing the random coil size by more effective charge screening of the backbone phosphates. As reported by Mfold[7e], calculated possible secondary structure for our leader mimic had low energy and few canonical base pairs. Future experiments were done with no added magnesium, as the TRAP protein is already known to bind effectively without added divalent metals[61]. This allowed us to disentangle any changes in apparent FRET from the condensing effect of magnesium. The measured $Mg_{1/2}$ concentration and lack of obvious cooperativity is similar to that measured for other unstructured RNAs of this size and consistent with nonspecific binding. The dynamics of individual traces for the 4 mM concentration seemed higher than at other concentrations. The resampled standard deviation of this concentration reflects this. This may be an experimental artifact, or an intermediate stabilization regime between two groups of undetected microstates.

4.4.2 Tryptophan saturated TRAP protein displays a Kd_{app} similar to other measurements at the single molecule level

The next property of the TRAP protein to be tested was the binding behaviour when the protein was completely saturated with tryptophan. This would allow us to probe the behaviour of fully activated protein binding to RNA without the added complexity of tryptophan also incompletely binding to the protein. In terms of what is already known about the protein, information for the saturated binding behaviour is plentiful, giving us a valid point of comparison to other researchers working with bulk techniques. This comparison would allow us to validate our method, and see the lowest protein concentration at which we could detect binding and check that this concentration was biologically relevant.
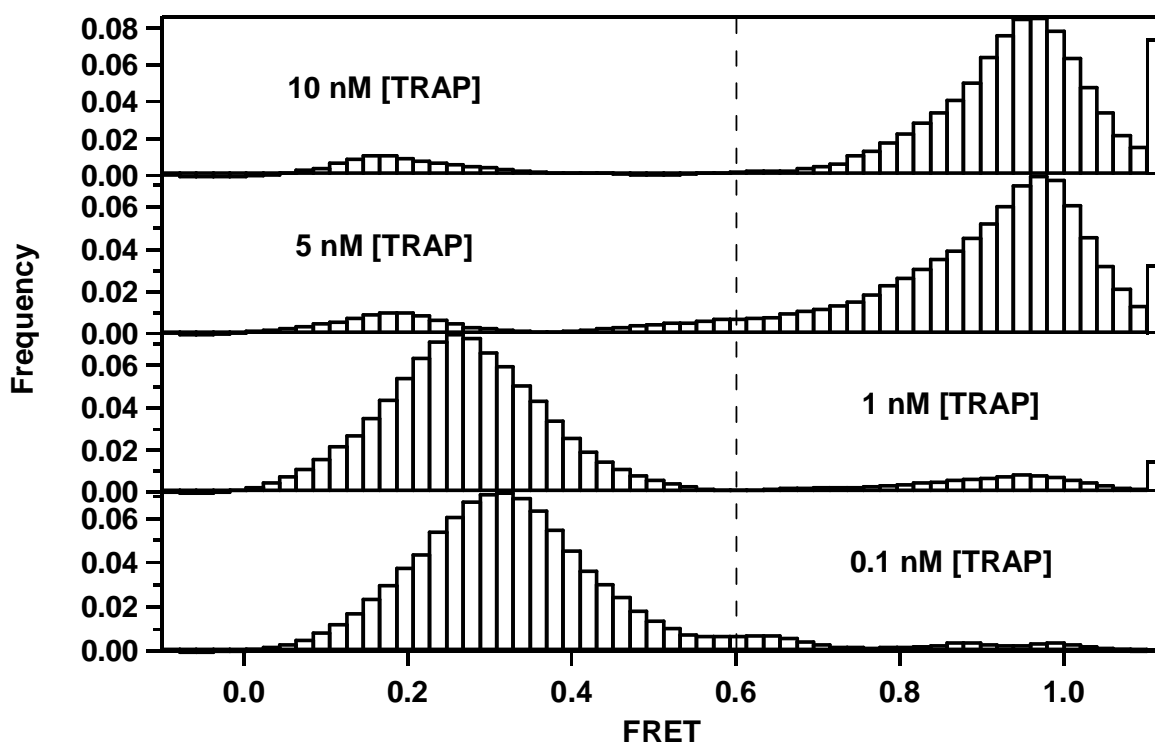
Figure 52: Histogram showing the behaviour of the RNA construct in the presence of the TRAP protein in a saturating concentration of tryptophan with varying protein concentrations. These experiments were conducted in 500 nM tryptophan, 50 mM MOPS, pH 7.5, 100 mM NaCl, PCD/PCA[59] as the oxygen scavenging system, and 2 mM Trolox[60] as a antioxidant and triplet state quencher. The protein concentrations are given in terms of the holo enzyme, not the subunit concentration (which would be eleven times greater).
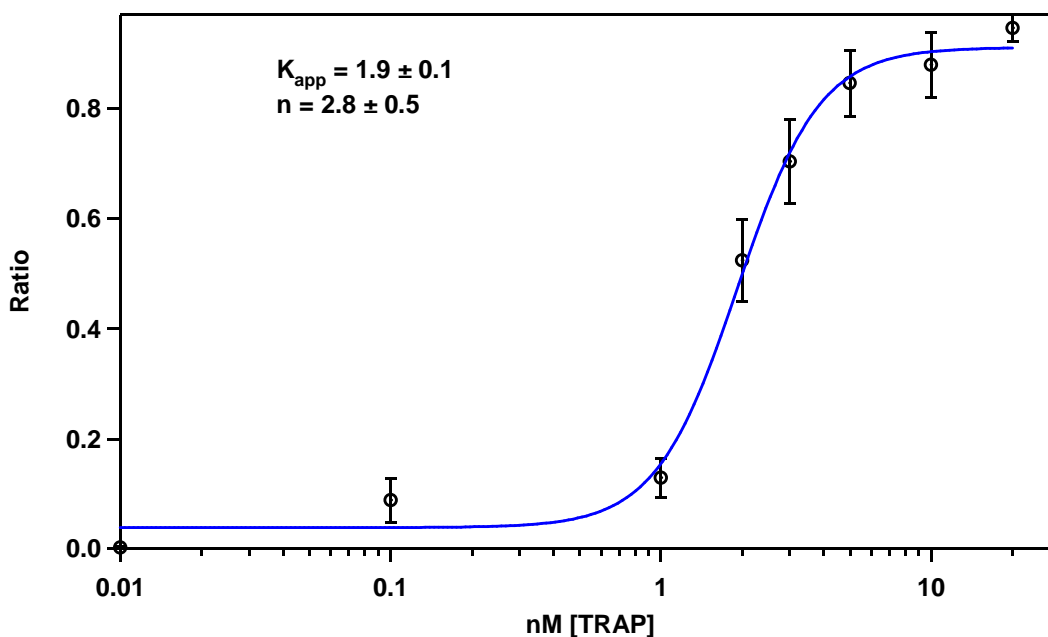
Figure 53: Binding curve for the RNA construct, in the presence of the TRAP protein, in a saturating concentration of tryptophan with varying protein concentrations. These experiments were conducted in 500 nM tryptophan, 50 mM MOPS, pH 7.5, 100 mM NaCl, PCD/PCA[59] as the oxygen scavenging system, and 2 mM Trolox[60] as a antioxidant and triplet state quencher. The protein concentrations are given in terms of the holo enzyme, not the subunit concentration (which would be eleven times greater).

Upon the addition of tryptophan saturated TRAP to the system, the response of the system changed in definite ways (see figures 52 and 53). We found two main apparent FRET states, a low fret state (approximately .3 FRET) and a high FRET state (approximately .9). We detected very little shifting from state in individual single molecule traces (3 molecules out of 500 analyzed). The amount of high FRET state found increased with increasing protein concentration. This is consistent with a fast on rate and a slow off rate. When fitted to the Hill equation we determined a Kd_{app} of 1.9 nM and a n of 2.8. Theses Kd_{app} and n values are consistent with values determined by other researchers[54]. The cooperativity is unexplained, but perhaps the TRAP protein is binding as a dimer. There is some evidence for dimerization from the crystal structure of Anston *et al.*[49] and the mass spectral work of McCammon *et al.*[62] and the work of Baumann *et*

al.[61]. However, gradient centrifugation, and gel electrophoresis data do not show the dimer forming in appreciable concentrations[61]. The actual state of TRAP oligomerization in the cell is then currently undecided. Our measured Kd_{app} and n values show that this single molecule system can accomplish monitoring of the TRAP protein binding to RNA, at concentrations that are close to cellular levels of the TRAP protein (estimated at 80 nM^{63}).

4.4.3 Increasing tryptophan levels alters the amount of bound protein in unexpected ways

Finally, the most exciting experiment to be done with the TRAP protein was to see the behaviour of the system in the natural regulatory context; namely what would happen to the binding properties of the protein as the external tryptophan concentration was altered. Previous work has found binding constants to be on the low nanomolar scale. We would like to see if our single molecule system reproduces these results, and if it will explain some of the other interesting properties of the TRAP protein. These include the seeming paradox of highly cooperative binding of the protein to RNA and the weakly cooperative binding of tryptophan to the protein and the requirement of 4-5 RNA binding motifs for protein binding to the RNA. We would also like to be able to detect the correct stoichiometry (*i.e.* one trap protein should bind eleven tryptophan molecules). Finally perhaps we can detect intermediate states or dynamics that have not previously been reported.
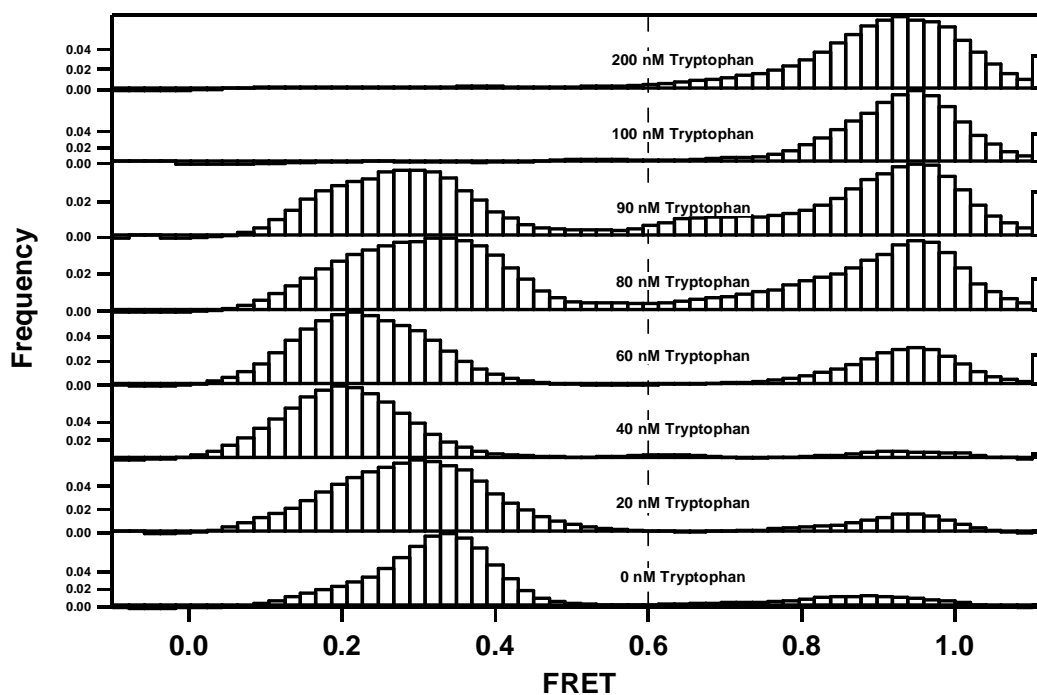
Figure 54: Histogram showing the behaviour of the RNA construct, in the presence of the TRAP protein, with different concentrations of tryptophan. These experiments were conducted in 50 mM MOPS, pH 7.5, 100 mM NaCl, PCD/PCA[59] as the oxygen scavenging system, and 2 mM Trolox[60] as a antioxidant and triplet state quencher. The protein concentration was 10 nM, in terms of the holo enzyme, not the subunit concentration (which would be eleven times greater).
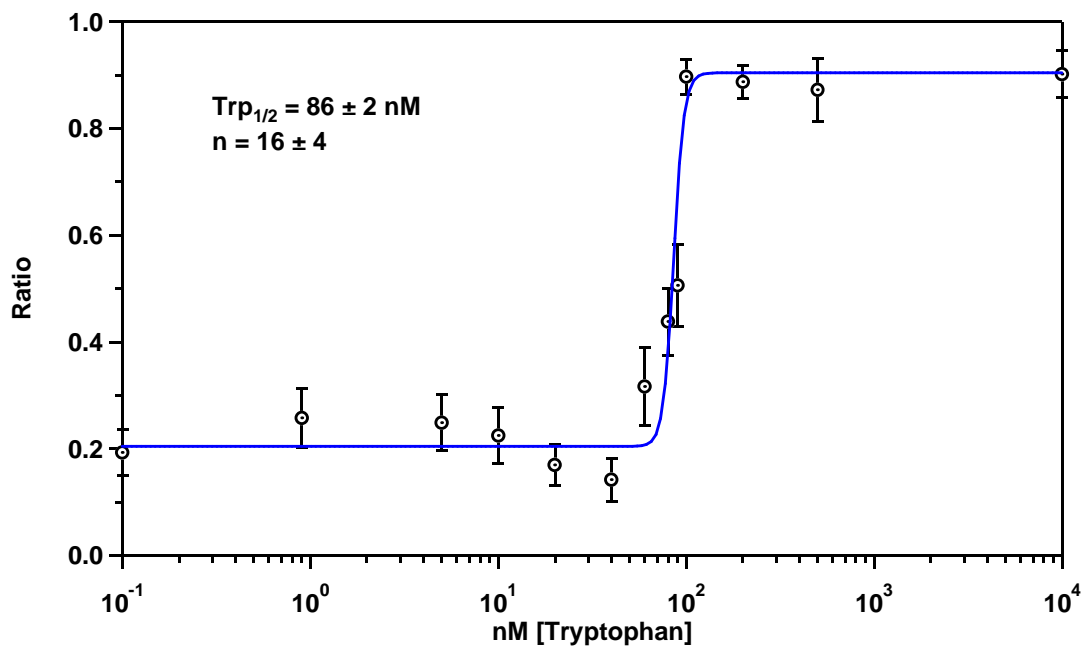
Figure 55: Binding curve for the RNA construct, in the presence of the TRAP protein, with different concentrations of tryptophan. The curve plotted is the best fit of the Hill equation. This is inappropriate to use in this case as there are at least two different equilibria (tryptophan binding to the TRAP protein, and the TRAP protein binding to RNA), but gives some rough values for comparison to other work. These experiments were conducted in 50 mM MOPS, pH 7.5, 100 mM NaCl, PCD/PCA[59] as the oxygen scavenging system, and 2 mM Trolox[60] as a antioxidant and triplet state quencher. The protein concentration was 10 nM, in terms of the holo enzyme, not the subunit concentration (which would be eleven times greater).

Inspecting the titration of fixed RNA and TRAP protein concentrations with tryptophan (figures 54 and 55), we detect the high FRET state (that we associate with the binding of the protein to the RNA), with no added tryptophan. This is in conflict with other authors which have reported that there is no appreciable binding in the absence of tryptophan[64]. We suggest that our single molecule assay may be selecting binding competent TRAP proteins from the pool of excess protein (the RNA is applied to the slide surface at 25 pM concentration; the protein concentration of the holo enzyme is 10 nM). This effect will be explained in more detail in the discussion section. After the initial binding event there is a small increase in the amount of binding from zero to one nM tryptophan. After this increase there is a drop in the amount bound, reaching a maximum at 40 nM. Then a very sharp increase is seen from 40nM with saturation at 100 nM tryptophan. This sharp increase is consistent with a cooperative binding event[65]. After this increase there may be a small drop in binding, from 100 nM to final saturation at 10000 nM tryptophan concentration, but this is not distinguishable from experimental error. This dual phase binding curve is unusual for regulatory proteins. As with the protein saturated with tryptophan, there were no detectable dynamics. Extending the observation time from 5.5 minutes to 30 minutes had no effect on this conclusion.

4.5 Discussion

4.5.1 Tryptophan Group Binding Model

To explain these interesting effects in the TRAP/RNA/tryptophan system, we offer two models. We suggest, as a first approximation, that each subunit of the TRAP protein has two conformational states, a tryptophan binding state, and a RNA binding state. These states are in equilibrium with each other. With no added tryptophan, some of the TRAP proteins have enough of the subunits in the RNA binding conformation to bind and give the basal level of binding seen.

Figure 56: Basal binding configuration for the group tryptophan binding model (red indicates the RNA binding conformation, green indicates the tryptophan binding conformation)

As low levels of tryptophan are added and bind, the binding event induces a conformational change in one or two subunits (as the tryptophan binds at the interface of two subunits[49]) to the RNA binding conformation. This change biases neighboring subunits away from the RNA binding conformation to the tryptophan binding conformation. This would result in a drop in the number of subunits in the RNA binding

conformation, for low tryptophan levels. Thus, small amounts of bound tryptophan would stabilize one subunit in an RNA binding conformation, but would destabilize that conformation in the neighbors. The net effect would be to reduce the total amount of bound protein. This biasing effect may or may not be detectably coupled to tryptophan cooperativity (we do not detect tryptophan binding; our assay is only sensitive to protein binding to RNA). Also, this effect on the neighboring subunits, may not be strongly coupled to overall tryptophan affinity, explaining the diversity in measured cooperativity in tryptophan binding to TRAP proteins from different species[52].

Figure 57: Low levels of tryptophan lead to a drop in binding for the group tryptophan binding model (red indicates the RNA binding conformation, green indicates the tryptophan binding conformation).

The large cooperative effect in binding RNA, as the tryptophan concentration increases, is due to the selection of particular states from the ensemble of possible

arrangements for bound tryptophan. It seems that for binding the TRAP protein requires something on the order of 4-6 of the RNA motifs (that the individual subunits recognize and bind) to bind effectively[54]. If instead, the arrangement of 4-6 tryptophan bound next to one another in the protein is the actual requirement, selection of the states where this is the case can result in the appearance of very cooperative effects. This sort of effect is seen in large ligand binding to one dimensional lattices, and our explanation draws much inspiration from the models proposed to explain these without allosteric conformational changes [66]. As an illustrative example, consider the case of 9 tryptophans bound to the 11mer TRAP protein. There are no possible arrangements of tryptophans in the protein where there are less than 5 tryptophans bound next to each other. The number of possible arrangements where this is the case, as the number tryptophans in the protein increases displays a trend that would result in cooperative binding. This selection of particular configurations would explain the lack of a decrease in free energy for greater numbers of RNA binding motifs past approximately 6[54]. As this state is sufficient for binding, increasing tryptophan concentration (and the number of tryptophans bound to the protein) would result in more proteins with the required distribution of tryptophan, and more bound (more than 6) tryptophan would have little effect on binding. Looking at this in a different way, as tryptophan binds to the protein it is more likely that another tryptophan binds to the protein in a site not adjacent to a previously bound tryptophan, until sufficient sites are filled (e.g. if one tryptophan is bound to the protein, there are two neighboring sites, but 7 sites that are not neighbors; it is more likely a new tryptophan binds to a non neighboring site if there are not strong cooperative effects).
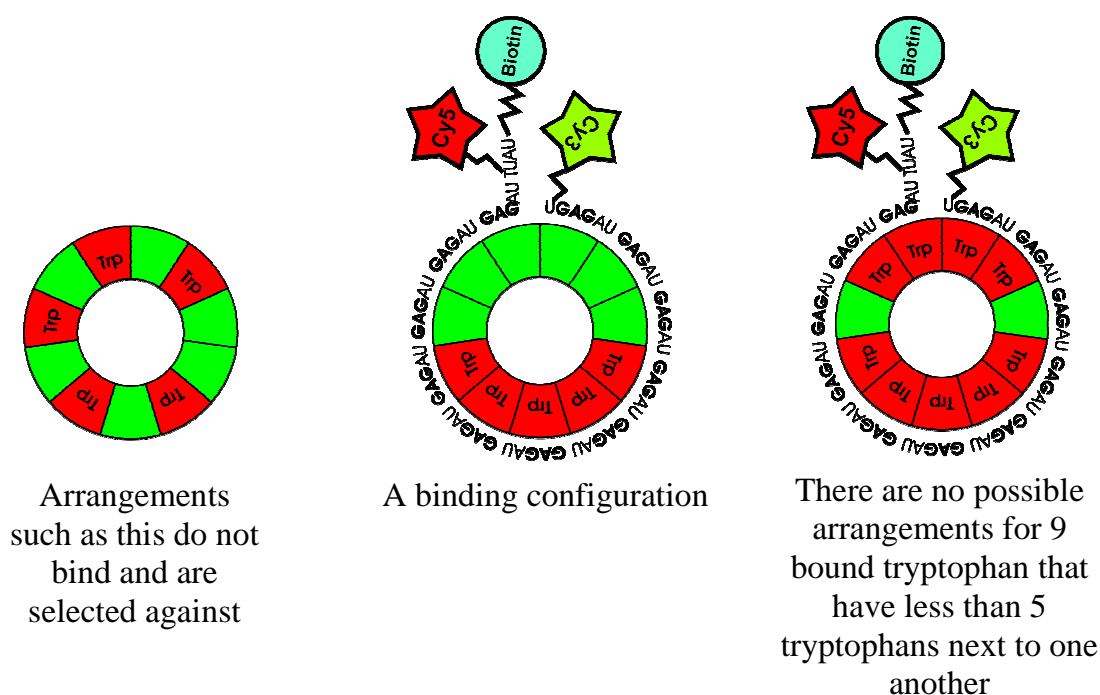
Figure 58: Origin of the cooperative effect in RNA/TRAP/Tryptophan system for the group tryptophan binding model (red indicates the RNA binding conformation, green indicates the tryptophan binding conformation).

The mechanism for this selection of 4-6 units grouped together is not determined, but could be due to the physics of initial interaction, as 4-6 subunits may be the number of subunits that initially interact and "stick" to the RNA. This amount of binding (approximately half the circumference of the protein) would be sufficient to bring the ends of the RNA together. We do not think the MWC model of a concerted conformational change describes this system (as is the case in hemoglobin), as NMR data does suggest this type of structural change[55]. Finally, the peak of low FRET state shifts with increasing tryptophan concentration. This could be due to transient interactions, with the RNA, that break up weak secondary structure or stretch out the initial random coil conformation.

No tryptophan

Increasing tryptophan concentration
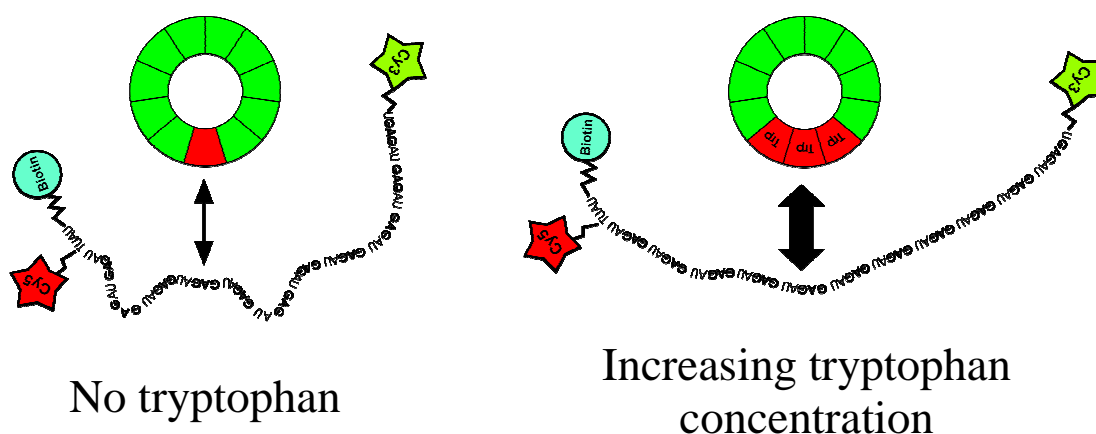
Figure 59: The origin of the shift in low FRET state for the group tryptophan binding model (red indicates the RNA binding conformation, green indicates the tryptophan binding conformation). Increasing interaction extends the random coil.

4.5.2 Distributed Tryptophan Binding Model

A different preference for the arrangement of bound tryptophan while preserving the statistical and selection aspects of the above model could also lead to cooperative effects. Li *et al.*[53c] have shown that synthetic trap protein assemblies composed of mutant nonbinding subunits and wild type binding competent subunits can have high affinity for RNA. They show there is evidence for strong binding to RNA with only one binding competent subunit. In view of this, our system could be monitoring the behaviour of the TRAP protein after it is already bound to the RNA. In this interpretation of our data, the protein is bound from the beginning of the titration; this would result from the high relative concentration of protein to RNA (the protein concentration in the titrations is 10 nM, the RNA concentration is 25 pM as applied to the single molecule slide; the effective concentration is much lower as only a portion of this is surface immobilized). This would offer enough different TRAP proteins (differing in the number of subunits in the RNA binding conformation) to the RNA to allow those with subunits in the RNA binding conformation to bind. If enough subunits are in the RNA binding conformation, the RNA

would wrap completely around the protein and give the detected high FRET state (for initial binding without tryptophan). Our assay would not detect protein that is bound to only a single subunit as this would not bring the ends of the RNA together.

Initially bound, but undetected protein

Increasing bound tryptophan extends the random coil
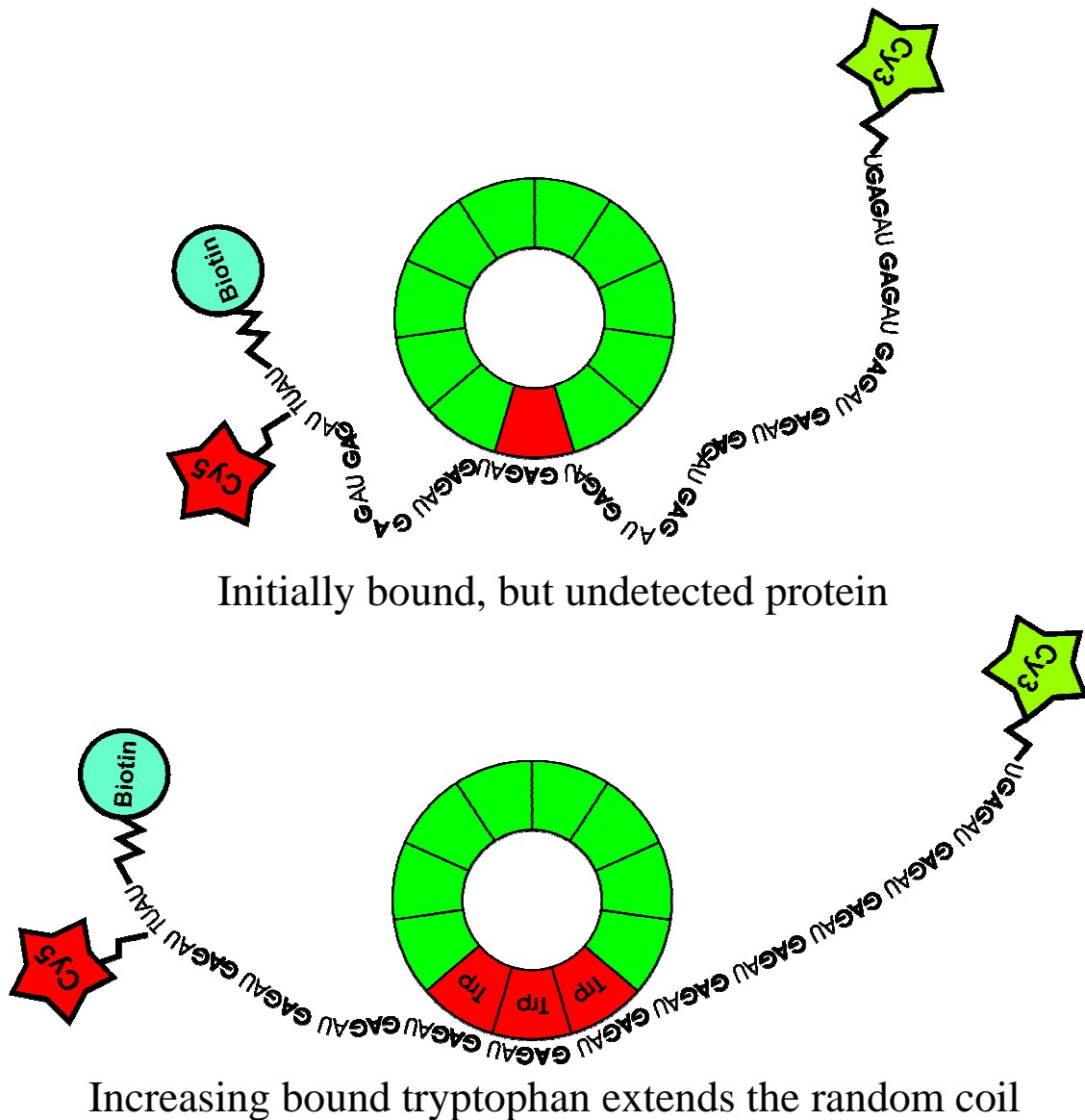
Figure 60: The cause of the low FRET peak shifting in the distributed binding model (red indicates the RNA binding conformation, green indicates the tryptophan binding conformation).

This single subunit binding could even extend the length of the RNA from the random coil length, or disrupt transient secondary structure (folding induced by secondary

structure formation would reduce the length of the RNA). This effect may be seen in the decrease of the FRET value for the low fret state in the 0 to 40 nM tryptophan. The high FRET state found initially (with no added tryptophan) is then due to a subset of the TRAP proteins that have enough subunits in the RNA binding conformation to bring the two ends of the RNA together.
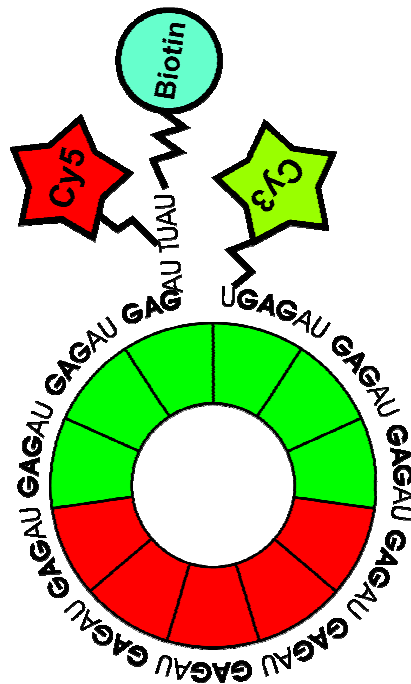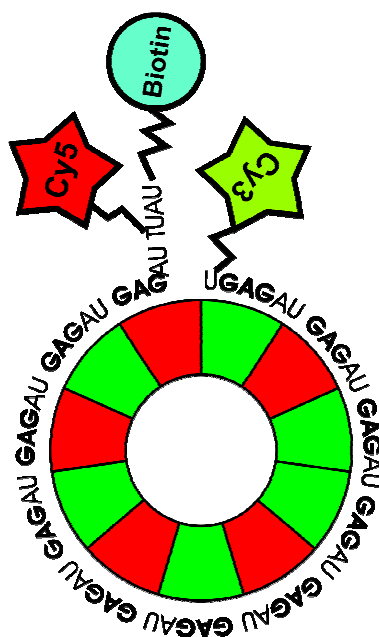

Figure 61: Basal binding configuration for the distributed tryptophan binding model (red indicates the RNA binding conformation, green indicates the tryptophan binding conformation).

As small amounts of tryptophan are added, a certain distribution of RNA binding conformations in the ring is required for "productive binding", *i.e.* binding that will bring RNA completely around the ring. Biologically, this would correspond to enough force to unfold the anti-terminator hairpin and allow the terminator hairpin to form. The drop in binding from one nM tryptophan to 40 nM tryptophan would be attributable to more RNA binding conformations near the initial site of binding. This would extend the RNA chain even further from the random coil length than a single bound subunit, changing the FRET value of the low FRET state. The reduction in the amount of high FRET state

would then be due to the effect of tryptophan bound in protein away from the initial binding site, which would bias unbound neighbors away from the RNA binding state, to levels below that seen with no tryptophan bound. This process is similar to the argument in the group binding model.

Initial bound state   Tryptophan binding induces a conformational change

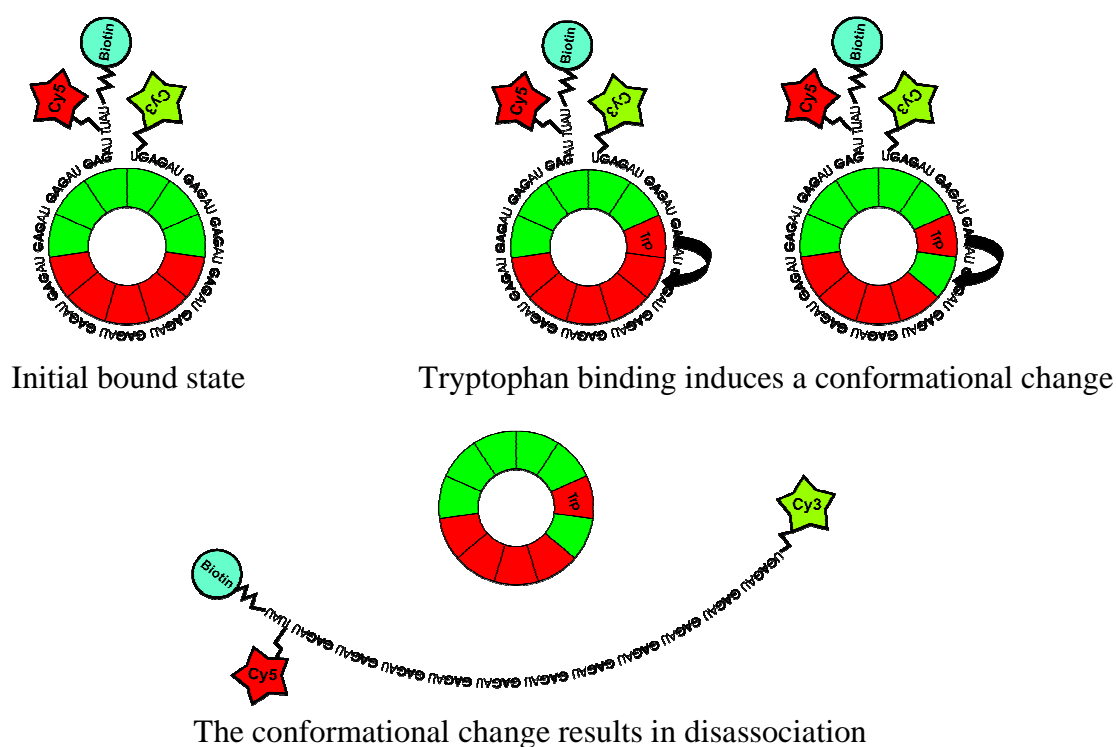The conformational change results in disassociation
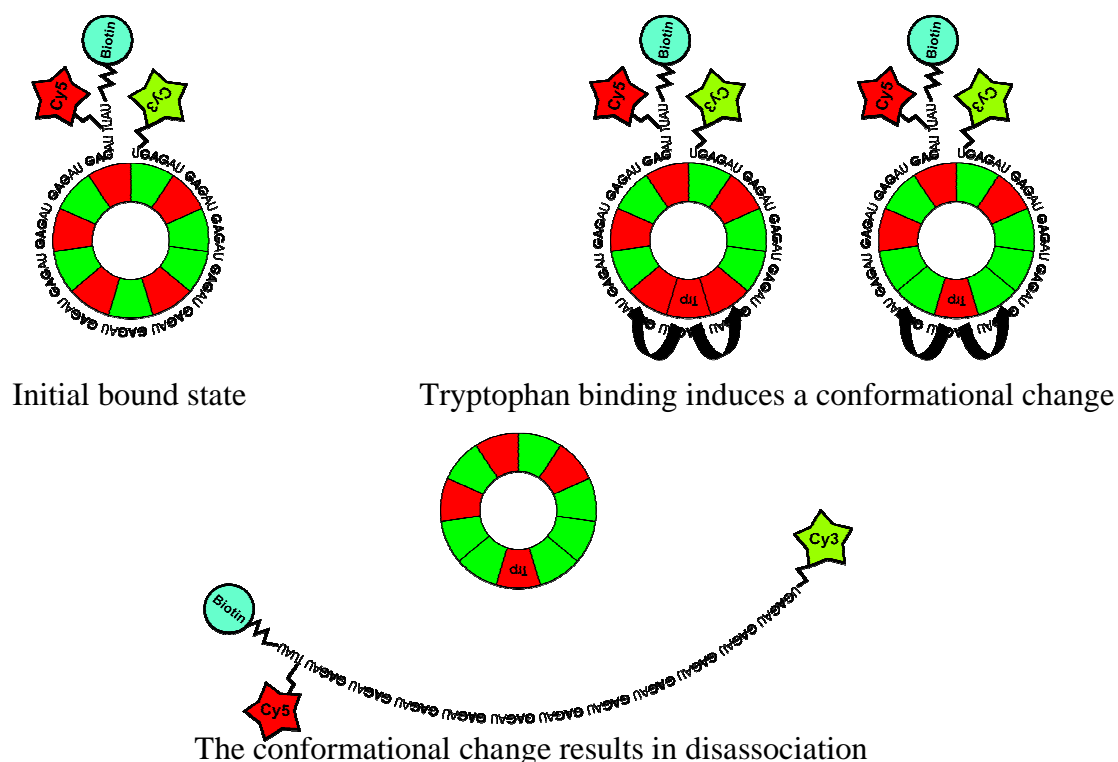
Figure 62: Low levels of tryptophan lead to a drop in binding for the distributed tryptophan binding model (red indicates the RNA binding conformation, green indicates the tryptophan binding conformation).

The strongly cooperative effect is, as in the group tryptophan binding model, due to a collapse in the number of states or arrangements that do not have tryptophan distributed around the ring. In essence, the states selected would be opposite to those in the first model. A slightly different way of restating this is that the RNA binds completely if there are no more than 2 subunits in the tryptophan binding conformation in a row.
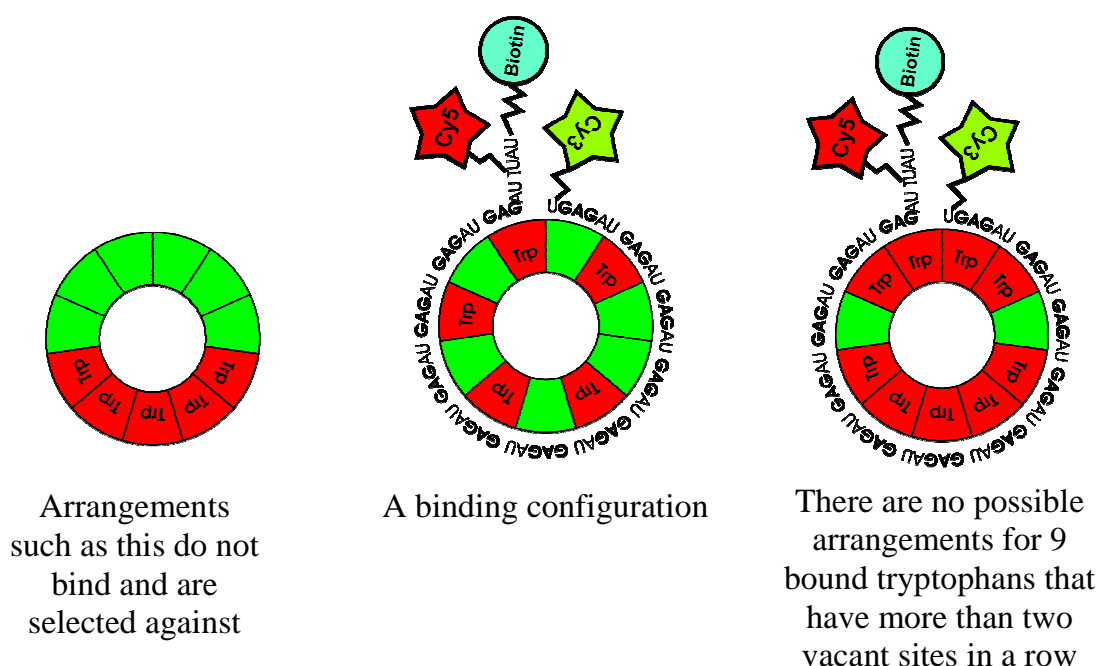
Figure 63: Origin of the cooperative effect in RNA/TRAP/Tryptophan system for the distributed tryptophan binding model (red indicates the RNA binding conformation, green indicates the tryptophan binding conformation).

An interesting feature of our modeling is that the small drop in binding from 100 nM tryptophan to 10000 nM tryptophan could be seen as a preference (over that of states that have neighboring RNA binding states) for states that have alternating RNA binding conformation and tryptophan binding conformations. If this drop is real, it could hint at a larger set of preferred conformations.

4.5.3 Mathematical Modeling

To determine if our qualitative models for binding were possible mathematically, a probabilistic equation reflecting the models was constructed. This mathematical model had a few basic constraints. Firstly, that each subunit could have only two states, an RNA binding state (this is a tryptophan bound state), and a tryptophan binding state (the subunit in this case would have no tryptophan bound). Secondly, that tryptophan binding

to a subunit induces a conformational switch from a tryptophan binding conformation to an RNA binding conformation. Said another way, this constraint meant that if tryptophan binds to a subunit, it instantaneously changes to the RNA binding conformation (this also means that if tryptophan leaves a subunit, it also undergoes a conformational shift back to the tryptophan binding conformation). Thirdly, that tryptophan binding to the protein could change the affinity of the subunit for RNA, but that RNA binding did not change the affinity for tryptophan. Finally, our fourth constraint was that the affinity of a subunit for RNA could depend on the conformation of the neighboring subunits (*i.e.* one subunit on either side). We decided to not include in the model conformations that are in the RNA binding state, but have no tryptophan bound, to simplify the expression somewhat. These constraints meant that there were six different probabilities for a subunit to bind RNA (depending on the conformation of the neighboring subunits, see figure 64) and two probabilities for tryptophan binding a subunit.
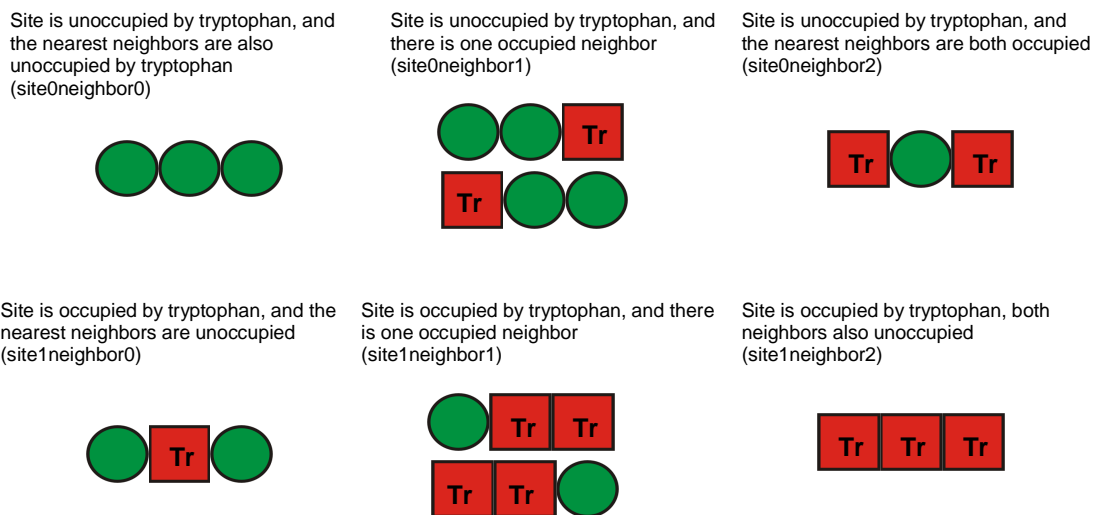
Figure 64: Pictorial representation of the configurations for a site and its nearest neighbors. There are two different degenerate arrangements for a site with one neighbor, and so are grouped together.

Since we are limiting our analysis to effects on RNA binding for subunits that are directly in contact, for each subunit the RNA binding probability is dependent on the conformation of the subunit and its two neighbors (one on each side, as there is a hole in the center of the 11mer TRAP holoenzyme; these are the only direct contacts for a particular subunit). There are then six different non redundant configurations for a site and the neighboring sites. These are: (1) The site is unoccupied (the site is in the tryptophan binding conformation) and the neighboring sites are unoccupied (in the tryptophan binding conformations); (2) the site is unoccupied and has one neighbor that is occupied (and therefore in the RNA binding conformation); (3) The site is unoccupied and both neighbors are occupied with tryptophan; (4) The site is occupied (and in the RNA binding conformation) and the neighboring sites are not occupied; (5) the site is occupied and one neighbor is occupied, and (6) the site is occupied and both neighbors are occupied (see figure 64). The two states that have one neighbor could have this neighbor positioned on either side, but as these states are degenerate they are considered together. The first tryptophan binding probability was dependent on the bulk concentration of tryptophan and the second was a parameter to account for tryptophan cooperativity (or anti-cooperativity). This second tryptophan probability was used if a neighboring subunit of the protein had already had tryptophan bound (*i.e.* this probability could be used to encourage or discourage tryptophan binding next to already bound tryptophan, mimicking cooperative binding). This allows separation of cooperativity in tryptophan binding from cooperativity in RNA binding. The probabilities for RNA binding for the different possibilities for nearest neighbor occupancy were set to be relative to the binding probability for the unoccupied site with no occupied neighbors

which was arbitrarily set to one. All possible arrangements of the subunits and neighbors in the context of the 11 subunit ring (2048 different configurations of subunits) were calculated and weighted by probabilities for the different neighbor possibilities using the computer program Mathematica[67]. The probabilities were introduced as changeable constants to allow fitting to experimental data. This was used to generate a final binding polynomial (or generally, a partition function). The Mathematica code and the final binding equation (binding to RNA) are shown in appendix B and an approximate pictorial representation is shown in figure 65.
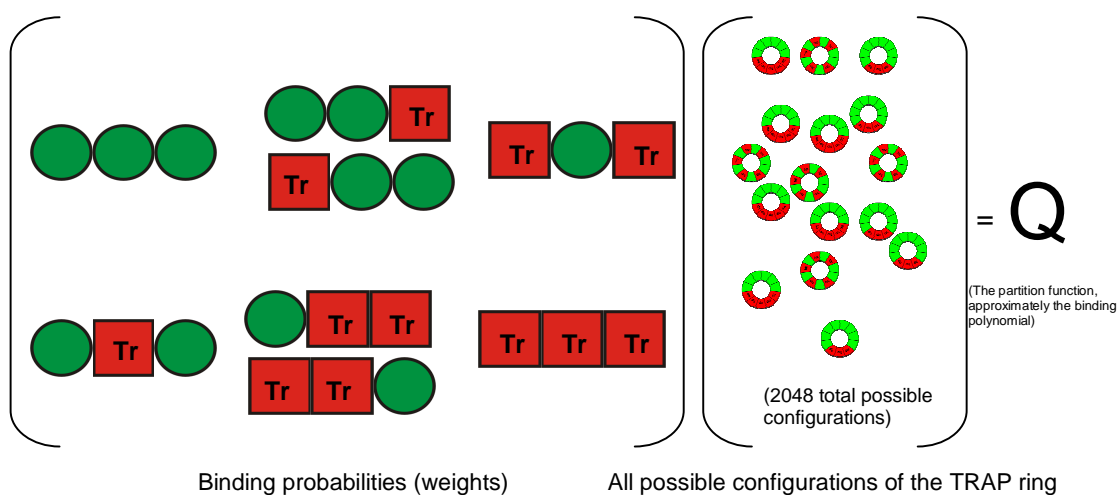
Figure 65: Approximate graphical representation of the Mathematica code used to generate the binding polynomial

The tryptophan binding probability in this equation shifts the participation of the different configurations (of the holoenzyme) from those with few tryptophans bound to those with many tryptophans bound. This means that that at higher concentrations it is more likely that a configuration with more bound tryptophan is present and is participating in binding. From this equation one can calculate a binding curve by plotting the amount

bound versus concentration (by using a correction constant for concentration to convert from probability to concentration,). The equations parameters were then be fitted to the measured data using the computer program IGOR pro [57] (Igor employs the Levenberg-Marquardt algorithm; this algorithm finds a local minimum for the fitted parameters.). While the many parameters in the model allow fitting to infinitely numerous curves, parameters can be found that embody the qualitative models (see figures 66 through 70).
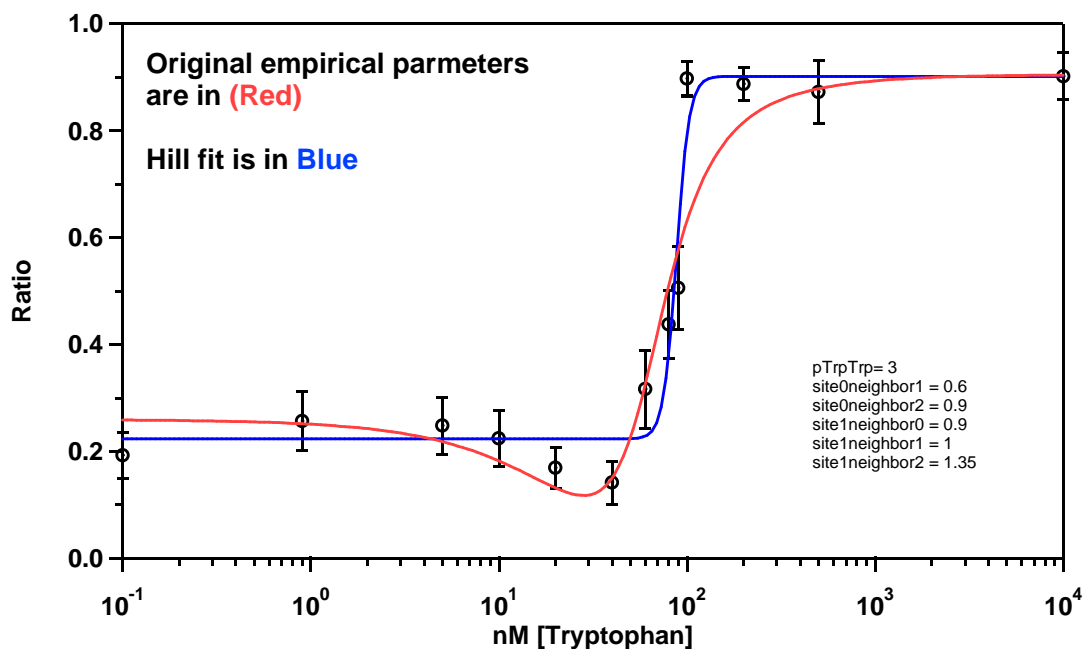
Figure 66: Statistical model plotted with the measured data from tryptophan titration. The red curve is that obtained substituting the statistical model with empirical parameters. The blue curve is the measured data fitted to the Hill model for cooperative binding. This model is plotted for qualitative comparison only as is not applicable to the TRAP/tryptophan/RNA system. The error bars are twice the standard error from resampling analysis (see Appendix A).
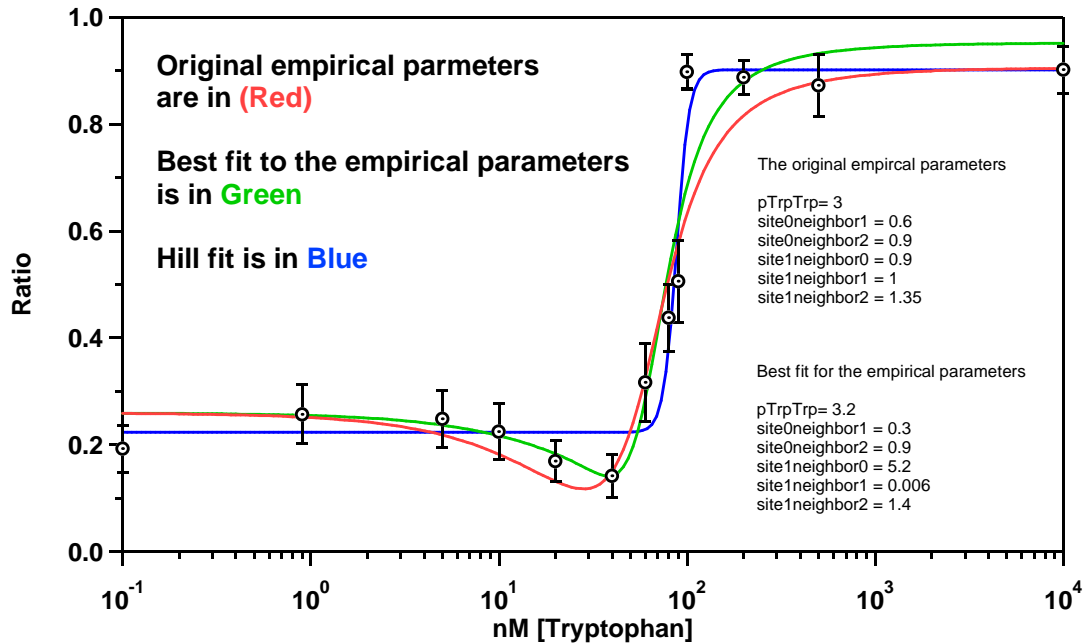
Figure 67: Statistical model plotted with the measured data from tryptophan titration. The red curve is that obtained by substituting empirical parameters into our statistical model. The blue curve is the experimental data fitted to the Hill model for cooperative binding. This model is plotted for qualitative comparison only as is not applicable to the TRAP/tryptophan/RNA system. The green curve is the best fit to the measured data when the fitting algorithm is started with the empirical parameters. The error bars are twice the standard error from resampling analysis (see Appendix A).
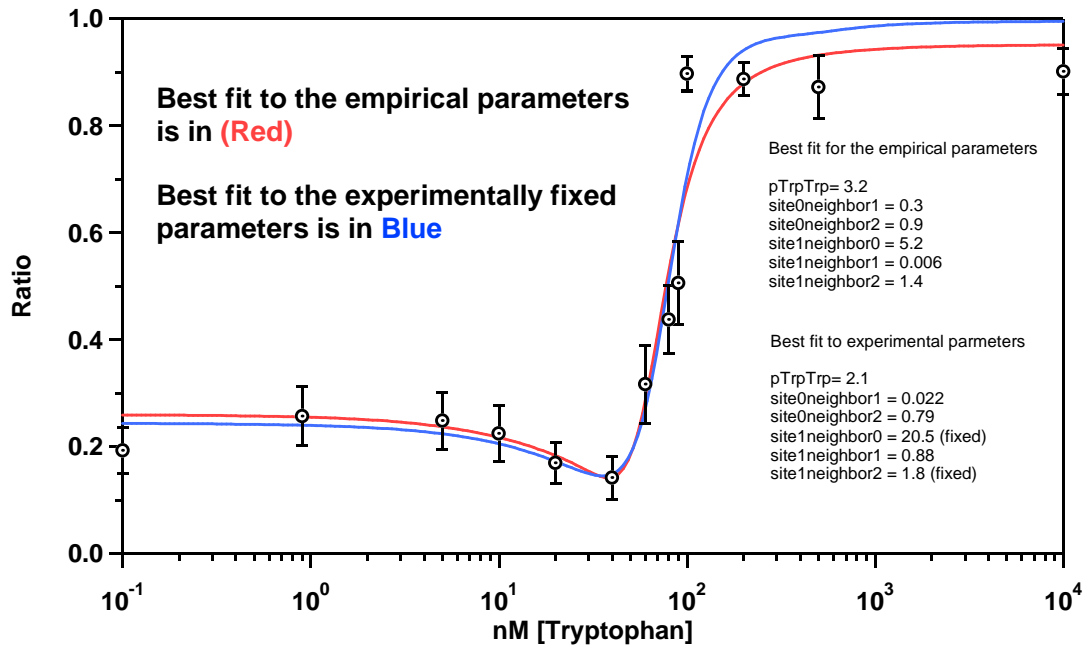
Figure 68: Statistical model plotted with the measured data from tryptophan titration. The blue curve is that obtained with experimentally fixed parameters. The red curve is the best fit curve generated by empirical parameters. The error bars are twice the standard error from resampling analysis (see Appendix A).

Best fit to the experimentally fixed parameters is in (Red)

Best to experimentally fixed parameters with low tryptophan cooperatity is in Blue

Best to experimentally fixed parameters with high tryptophan cooperatity is in Green

Best fit to experimental parmeters

pTrpTrp= 2.1
site0neighbor1 = 0.022
site0neighbor2 = 0.79
site1neighbor0 = 20.5 (fixed)
site1neighbor1 = 0.88
site1neighbor2 = 1.8 (fixed)

Best fit to experimental parmeters
with low tryptophan cooperativity

pTrpTrp= 0.5 (fixed)
site0neighbor1 = 0.01
site0neighbor2 = 1.5
site1neighbor0 = 20.5 (fixed)
site1neighbor1 = 0.27
site1neighbor2 = 1.8 (fixed)

Best fit to experimental parmeters
with low tryptophan cooperativity

pTrpTrp= 4.2 (fixed)
site0neighbor1 = 0.13
site0neighbor2 = 0.36
site1neighbor0 = 20.5 (fixed)
site1neighbor1 = 1.7
site1neighbor2 = 1.8 (fixed)

Figure 69: Statistical model plotted with the measured data from tryptophan titration. The red curve is that obtained with experimentally fixed parameters. The blue curve is generated using the estimated measured parameters, and fixing the tryptophan cooperativity (pTrpTrp) to an arbitrary low value. The green curve is generated using the estimated measured parameters, and fixing the tryptophan cooperativity (pTrpTrp) to an arbitrary high value. The error bars are twice the standard error from resampling analysis (see Appendix A).
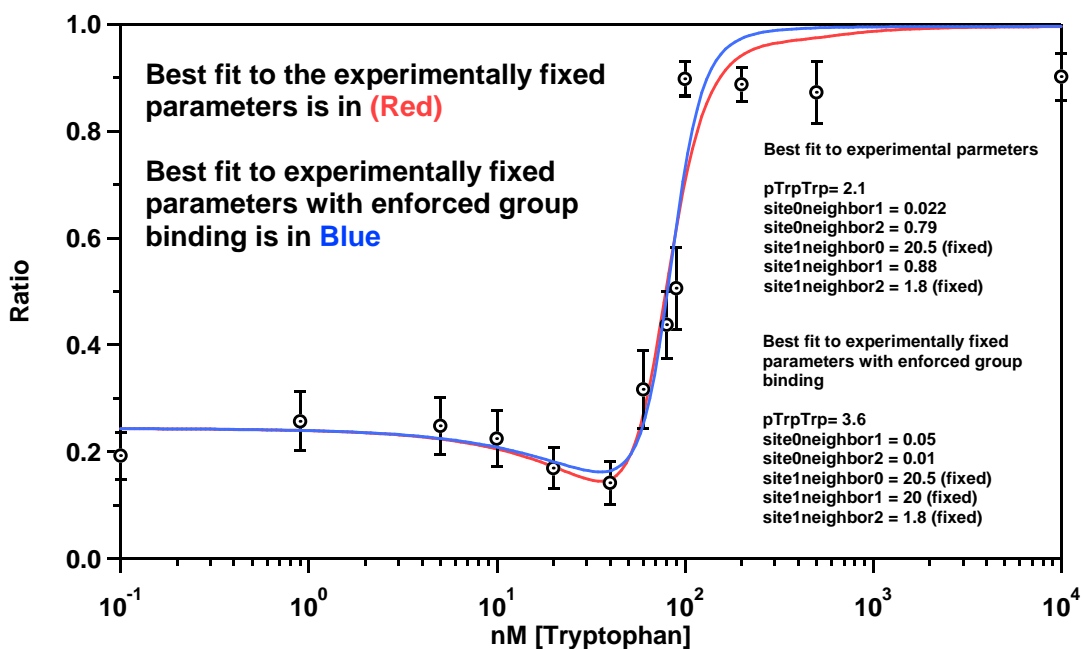
Figure 70: Statistical model plotted with the measured data from tryptophan titration. The red curve is the best fit obtained with experimentally fixed parameters. The blue curve is generated using the estimated measured parameters, and fixing the binding probability for an occupied site with one neighbor (site1neighbor1) to an arbitrary high value. The error bars are twice the standard error from resampling analysis (see Appendix A).

These model parameters indicate probabilities (relative to a site that is unbound with no bound neighbors) that a site with its nearest neighbors in a defined configuration will bind RNA, and so can be easily interpreted in the context of the models. Initially, we empirically assigned parameters to give a curve that mirrored the experimental data. These empirical parameters show that a process that is cooperative in tryptophan binding (the probability for tryptophan binding next to another tryptophan binding is greater than one) can mimic our data. The probability for binding to an unoccupied site with one or two neighbors is less than that for an unbound site with no neighbors. Also, the probability for binding to a site which is in the RNA binding configuration (tryptophan bound) with no neighbors bound is less than that for the completely unbound situation. This could because we are setting our probabilities relative to a state that can bind RNA

(the unoccupied site with no occupied neighbors is in conformational equilibrium with the RNA binding conformation, we see this as a basal level of RNA binding). This is in line with our model, in that we propose that small amounts of bound tryptophan can produce a drop in affinity from that seen with the tryptophan free form of the protein (creating more configurations where states exist with neighbors). The numbers of states found with only a few bound tryptophans would be heavily weighted towards those without neighbors, as each bound tryptophan without neighbors creates two neighbors with a bound neighbor. Interestingly, the first probability that is greater than that of the completely unbound configuration is that for the occupied state with fully occupied neighbor sites. Taken all together, these weights correspond to a preference (in RNA binding) for tryptophan bound in groups, as proposed in the group binding model. This group binding preference in combination with positive tryptophan cooperativity can result in the highly cooperative RNA binding process in the TRAP system.

To refine our empirically assigned parameters we imported our binding equation to the computer program IGOR pro[57] to fit the parameters to our measured data. After fitting by IGOR pro[57], two of our empirically defined parameters were found to have changed. The probability to bind RNA by an occupied site with no neighbors (site1neighbor0) increased by more than five times, and that of a bound site with one neighbor (site1neighbor1) dropped dramatically. These changes did not however have a large effect on the calculated binding curve. A reasonable interpretation for this is that with the number of parameters involved in the model, the calculated binding curve is only weakly dependent on any of them and that some parameters can substitute for others in terms of replicating our experimental data (cross correlation, and covariance, *e.g.* the

parameters do not have independent effects on the shape of the function). An obvious example of this effect is that the tryptophan cooperativity parameter and the site neighbor parameters are linked. In terms of final RNA binding, a high likelihood for tryptophan to bind next to another tryptophan can be offset by a low likelihood for a site with bound neighbors to bind RNA or *vice versa* (this effect can be seen in figures 69 and 70). Keeping these caveats in mind, a possible physical rationale for the fitted parameters can be attempted. This weighting for the parameters indicates that combination of the RNA binding conformation alone (with no neighbors) binding RNA strongly, combined with weak binding for an occupied site can also mimic our data. In terms of our models, this would correspond with distributed binding, where there is a preference for a staggered arrangement of bound tryptophan. The other parameters show disinclination in binding to unoccupied sites, consistent with the distributed model. These two parameter sets show that both of our proposed models are qualitatively consistent with our mathematical model.

This model is weak in that there are very many parameters (eight in total). This flexibility in parameter space allows many other qualitative models to also be interpretable in the framework of our mathematical model. In order to reduce the number of flexible parameters, and therefore reduce this flexibility in fitting, we made an attempt to assign some of the parameters to measured values for the TRAP/RNA/tryptophan system. To make this guess at the real physical values, we employed the work of Li *et al* [53c]. This was possible as these workers measured Kd values for holo TRAP proteins (the 11mer assembly) that were composed of mixtures of mutant and wild-type subunits. The mutant subunits were selected to be RNA binding incompetent. The final composition of

the assembled TRAP holoenzymes was controlled statistically (by the initial concentrations of wild-type and mutant subunits mixed together and allowed to form the holoenzyme). The authors claim that for some percentages of mixture (low amounts of wild-type subunit) that there is high probability that only one subunit in the ring is wild-type. For increasing percentages of wild-type in the initial mixture the number of subunits found increased. Also, these authors were also able to determine a Kd for a holo TRAP protein made of their binding incompetent subunits, saturated with tryptophan[68]. These Kd values give us some general reference points for the binding affinity for the unbound case with no neighbors (site0neighbor0), that of a site that is bound with no neighbors (site1neighbor0), and for a site that is bound with two neighbors (site1neighbor2). Substituting these derived parameters into our binding equation and fitting, we can also find curves that seem to fit the data (see figures 69-70). For these parameters we can also find parameter sets that correspond to both of our proposed models. Also, there are values for the parameters that can support anti-cooperative to highly cooperative tryptophan binding. This flexibility in parameters demonstrates that even with two of the parameters fixed, those remaining are sufficient for a multitude of different models. Setting this overabundance of parameters aside, this could be an explanation for the reported diversity in tryptophan cooperativity for TRAP proteins from different species. The underlying system creating the cooperative effect in binding may not be very sensitive to tryptophan cooperativity, and that the other binding parameters can compensate for changes. The important parameters may be the relative binding probabilities between the postulated states. In both the empirically derived parameters, and the ones guessed from measured parameters the binding probability for the unbound states, (excepting the

unbound state with no bound neighbors, site0neighbor0) are generally low. The probabilities for states have bound tryptophan are higher than those without. This difference in probabilities between bound and unbound states seems to be sufficient to reproduce our experimental data (roughly two to twenty times greater). These probabilities are consistent with expectations from our qualitative models, but unfortunately not sufficient to rule out either of them as possibilities. We can say that a mathematical model based on statistical thermodynamics that reflects our qualitative models is reasonable. If we can obtain more fixed values for the parameters in the model, we may be able to come to solid conclusions about its validity as the currently fixed parameters are not sufficient.
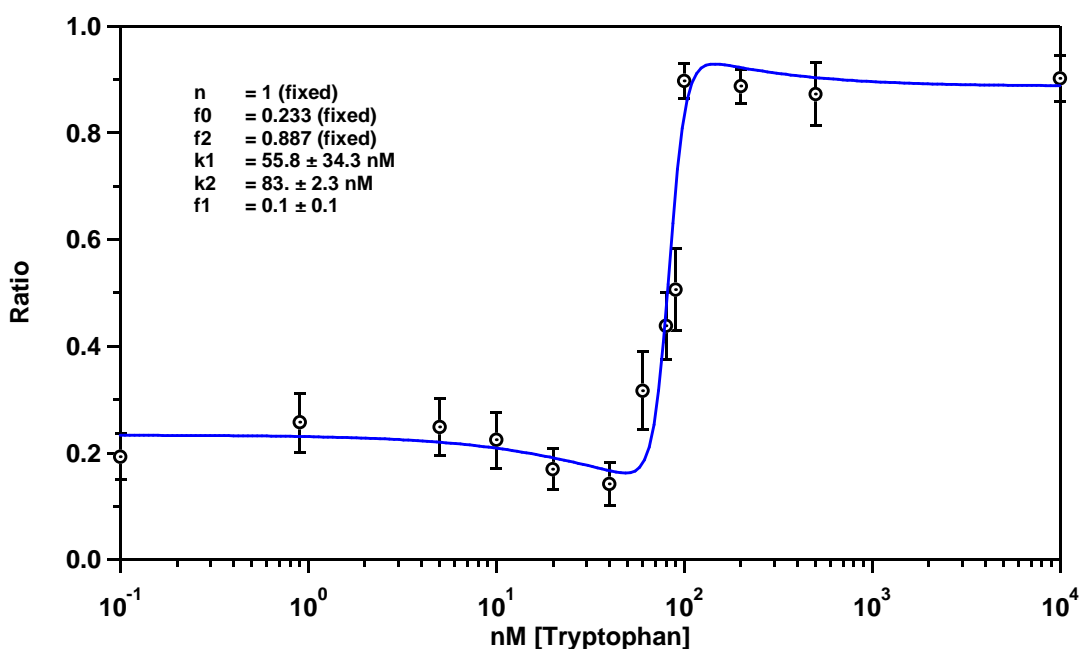
Figure 71: Data fitting with our minimal model. The cooperativity parameter was fixed to one for the first transition. The initial ratio was fixed to the average value for the first three points and final ratio was fixed to the average value for the last three points in the titration. The fitted parameters are inset in the graph.

$$fraction bound = f0 + (f1 - f0)\frac{[Tryptophan]^n}{[Tryptophan]^n + k1^n} + (f2 - f1)\frac{[Tryptophan]^{11-n}}{[Tryptophan]^{11-n} + k2^{11-n}}$$

Figure 72: The minimal binding equation. It consists of an initial bound ratio (f0), a cooperative transition to an intermediate bound state (f1), and a cooperative transition to a final fully bound state (f2).

Due to the problems inherent in a complicated model with many parameters we also explored a minimal model. This was to simplify the mathematics to the least complicated to adequately fit the data. This model consists of three characteristic ratio parameters and two cooperative processes. This is a minimal interpretation of our more complicated qualitative and statistical models. In short, we assume an initial state for the apo-protein that has some propensity to bind RNA (this would correspond to the initial random fluctuation of the subunits into a binding conformation). This initial state gives a characteristic ratio that corresponds to the initially bound parameter f0. After this initial state, as more tryptophan binds, there is an intermediate state, which is an aggregate of many microstates that binds less RNA (this would be the states in which some of subunits have biased their neighbors to tryptophan binding conformations, away from the RNA binding conformation). This corresponds to the second ratio parameter f1. Finally there is a highly cooperative transition to the final tryptophan saturated state that corresponds to the parameter f2. This model can also adequately fit the data. It gives parameters near to those measured by other groups except for that of the initial binding affinity. This may be an effect of the surface immobilization. This model shows that even without resort to an arguably over parameterized model, our basic qualitative model reflects the data.

There are many interesting questions still left open in the TRAP regulatory system. Why does the system need regulation at both the translational and transcriptional levels? With the presence of regulation at these levels, why is the system also coupled to

charged tRNA levels? How does a non cooperative process result in cooperative binding? Our models offer an explanatory framework for these processes, as well a suggesting some interesting evolutionary themes. The TRAP protein seems to have a basal level of binding to the leader element. This may be due to unrestrained transcription and translation resulting in too much tryptophan. This low level of repression may not be tolerable when the cell is stressed to the point where charged tRNA^{trp} levels have dropped, stopping protein synthesis. The anti-TRAP protein is then needed to completely shut down TRAP binding, allowing quick recovery. Regulation of tryptophan system at the transcriptional and translational level may be an effect of the stochastic nature of state selection. If a particular protein bound to the mRNA, or in the local neighborhood, never happens to display the required states, or in the appropriate transcriptional timeframe, a backup method is needed; explaining the presence of translational regulation in the same protein. Finally, our models suggest that purely statistical and physical considerations can result in very cooperative effects. These types of effects have been seen before in large ligands binding to one dimensional lattices (e.g. RNA and DNA[66]) and postulated to occur in regulatory protein binding to nucleosomal DNA[69]. The drop in binding at approximately 40 nM tryptophan has three biological effects. One, this results in a sharper transition to full binding, giving effectively switch like behaviour when toxic levels of tryptophan have the possibility of being produced. Two, this reduction in binding would allow rapid tryptophan production in times of cellular plenty, for quick growth and division. Three, combined with the higher basal level of binding, allows another regulatory regime. This regime would slow production of tryptophan with low non optimum cellular conditions, allowing redirection of cellular efforts the synthesis of

more common amino acids. As tryptophan is a comparatively infrequently used amino acid this would allow near normal function in times of moderate stress. When stressed to the point that there are low levels of charged tRNA^{trp} that interfere with normal protein synthesis, the anti-TRAP protein takes over, allowing unrestrained emergency production of tryptophan. As with our postulated statistical cooperativity this subtle three response system may be useful for other cellular processes and should be looked for.

4.6 Evolutionary and Nanotechnological Speculation

We posit that these types of effects can offer an evolutionary path to true allosteric cooperativity while still allowing a functional regulatory system. In an ancestral organism, many of the regulatory processes may have been tuned by the number of subunits in a regulatory protein, or through physical means the type or number of states in the ensemble of bound states (this may tune the affinity or the apparent cooperativity). The number of subunits could be increased or decreased by expression level or repetition of genes. After this a slower process of mutation of individual residues in the protein could slowly create true allosteric changes in the multimeric assembly, and efficiency in binding.
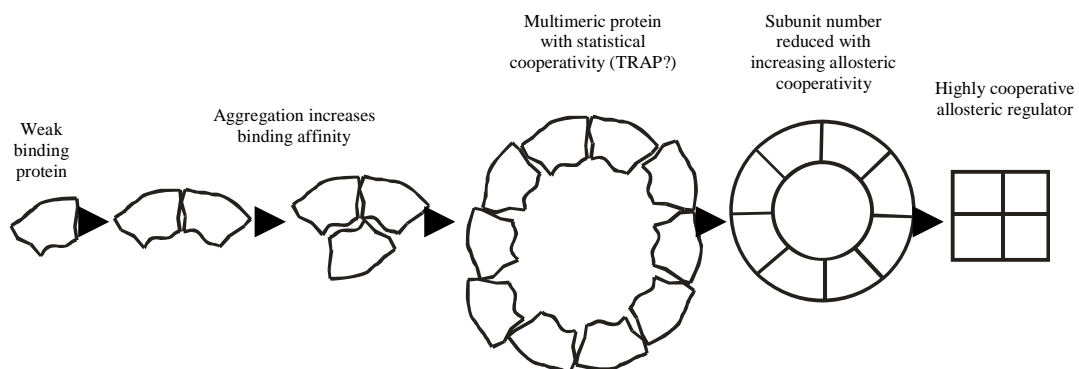
Figure 73: Proposed evolution of an allosteric protein.

Also if the protein has multiple instances of its gene, this allows segment swapping, duplication, and mutation in the whole structure without loosing all function. As the allosteric effects become greater, the need for many subunits becomes less, allowing their elimination and metabolic savings. This process would give an organism a functional regulatory system while the elaborate interdependent interactions required for concerted allosteric changes could slowly evolve. Seen in this light, the TRAP protein could be a molecular fossil displaying a once much more prevalent binding and regulatory plan. The TRAP protein may have been caught in a highly oligimeric state because of the number of the number of places in which it conducts a regulatory role. In the case of the TRAP protein, these multiple sites would have to mutate in a concerted manner to preserve the multiple functions. For others, this may explain the presence of the large number of multimeric allosteric proteins, as leftovers from a non-allosteric past with greater numbers of subunits. Interestingly, if this hypothesis is correct, this might allow the attempt at mutational archeology. It might be possible to mutate current multimeric allosteric proteins to preserve much of their regulatory or catalytic function, by increasing the number of subunits, and reducing the amount of allosteric cooperativity. There may even be a deducible ordered pathway of single mutations, or contiguous chunks of protein (swapped during recombination). Non-homologous recombination has been proposed as an efficient method of generating folded functional proteins[70]. When this process is with mimicked experimentally with exons to generate proteins, many of the proteins found are oligimers, sometimes of high order (stable tetramers have been found[70c]). This could also be seen as leftovers from a less allosteric and more oligimeric past, *i.e.* the mutation process has taken the first backward step by using pieces that have not been selected as

strongly for improved function or by eliminating improvements that reduced oligomerization. There is also evidence for the first step, that random protein sequences have a preference for oligomerization[71]. When random peptide libraries are constructed with a reduced amino acid set (chosen in such a way as to mimic historical conditions) the produced, soluble, proteins have a tendency to be multimers[71]. This retro-mutational process could allow stepping backwards to a hypothetical progenitor with only a single subunit, and perhaps multiple binding affinities. This type of analysis might also show a new way for proteins to be related to one another; there may be families related by this process that are not obvious from the sequence or structure. Oligomerization during protein mutation has been ignored in many experiments, as this is often seen as undesirable and ignored (i.e. the original protein is a monomer, mutations that encourage oligomerization are undesirable as they causing too much structural change to be useful for the rationale of the study), and so this theme may have been ignored. With modern mutational techniques, we suggest this retro-mutation is a feasible project. We recommend this sort of effort as a way to peer into the past of life, and illuminate central themes of cellular organization, and to see the progress from simple to more complex. It may be possible to reverse engineer to ancestral proteins that possess only one unit, which then evolved to multimers, then branched to other functions, and became allosteric and specific. A summary of the hypothesis from progenitor protein to the current day might look something like this: an ancestral protein randomly has some sort of effect on catalysis or RNA or DNA expression. It is poorly folded and not very functional. Nonspecific aggregation binds two or more of the proteins together, and function improves. Evolution selects for improved function and improved association. The initial

progenitor protein may start with many functions, or the mutational process may create them. The selection for multiple copies is much more rapid than point mutations (through recombination). A stable number of subunits are found. Point mutations select for improved function. This improvement in function may be seen in allosteric changes or improved binding or catalytic pockets. This improvement in function allows appropriate function with fewer numbers of subunits. Lack of evolutionary pressure for larger numbers of subunits allows the protein to eliminate subunits, eventually giving highly allosteric proteins with few identical subunits. This pathway, if correct, would suggest metrics for back mutation towards ancestral proteins, *i.e.* a decrease in allosteric cooperativity, an increase in the number of subunits, with general preservation of the function in question. We strongly encourage other research groups to start mutationally (or perhaps structurally/computationally) digging for molecular artifacts.

Aside from the fundamental interest in cooperativity and regulation, these models have implications for the design and function of nanostructures. Our model suggests that mechanical action (and the tuning of the action to respond to an external stimulus) can be performed by assemblies of identical subunits that do not interact with each other in a very integrated fashion (*e.g.* allosteric cooperativity), and that these assemblies could be a stepping stone to more efficient function. The amount of mechanical force or the magnitude and timing of the response may be able to be tuned by the amount of oligomerization, or interactions of a single subunit. This would allow an easier design path as the intricate intermeshing and coordination of an allosteric conformational change would not have to be created for useful function. Also, our evolutionary speculation could offer another path for the design of an optimized nanomachine. A search could be

conducted in the exon space (or through random sequence space) of a genome for small peptide units that accomplish a desired task. These could then be mutated and selected to find more efficient derivatives, and a signal for a productive mutational pathway would be the increase in oligomerization. After this oligomerization, the next signal would be a decrease in the number of subunits and an increase in function.

4.7 Conclusions

We have developed a single molecule assay for the tryptophan/TRAP/RNA system that mimics cellular changes and agrees well with previous work. Several novel models are offered to explain the behaviour of the system, with an underlying statistical interpretation of cooperative effects tying the models together. The evolutionary implications of the models have been explored. A broad suggestion for a different design paradigm for nanostructures based on the insights from the TRAP system has been put forth. As seen, our experiences with the TRAP system have been quite productive and leave open many avenues for others to explore.

5. Final Conclusions and Future Experiments

These three projects in self assembling nucleic acid design and synthesis have left many opportunities for others to build on them. The first project has been much built upon by DNA Software Inc. and will be soon released as a full fledged program, RNA-123, to simulate and predict the 3D structure of nucleic acids. This aims to be a transformational technology in terms of a researcher's ability to make reasonable designs and predictions of structure *de novo* from sequence information, or by leveraging homology information from related structures.

The second project designing and synthesizing a DNA nanostructure leaves the most to be done. Our work indicates that we designed and synthesized structures of the correct molecular weight and size. Much more remains to be done in insuring that the correct geometry was achieved. Also, as only one designed sequence (and correspondingly, structure) was tried the methodology has not been fully vetted in terms of general application. We suggest that future researcher use our methods but try many more different sequences and employ more effectively molecular imaging techniques. This may allow our initial goals to be realized, the easy synthesis of a complicated structure, in one pot, with minimal purification.

Our last project with the TRAP protein was our most successful and has the most potential for fundamental advancement in of the basic principles of nucleic acid design and the biophysics of a dynamic nucleic acid regulatory system. This sort of statistical cooperative process in binding for regulation of a cellular process that we propose has not before been used to explain how a single ligand can act as an antagonist and an agonist at different concentrations. If this hypothesis proves correct, it is probably a larger theme in

terms of regulation in biological systems, and should be looked for in other organisms and regulatory systems. This then can be used as biological inspiration for nanomachine design and construction. This also argues for the employment of single molecule techniques on a broad basis in biochemistry to uncover more previously unknown processes.

In terms of the TRAP protein system there are interesting questions that the present work does not address. One question is the role of physiological magnesium concentrations. We have had an interesting hint of this in our magnesium titration at 4 mM, in that our resampled standard deviation is much larger than with the other points. This could be experimental error, but also might indicate different dynamic processes occur at this concentration, and this may change the behaviour of the regulatory system. The experiment should be repeated with differing magnesium concentrations to answer this question. Our experiments were done with only one type of TRAP protein (from *B. stearothermophilus*), there are many other known TRAP proteins from different *Bacilli* species. We argue that this statistical cooperativity may be a larger theme in binding, so can we the same results in related TRAP proteins? We propose the role of the anti-TRAP protein is to shut down the baseline down regulation of the *trp* operon (the basal level of binding we detected in the absence of tryptophan). If similar experiments are conducted in the presence of the anti-TRAP protein will we see the absence of this basal binding? Finally, the 5' hairpin of the leader element of the *trp* operon has been postulated to play a role in helping the TRAP protein more effectively bind to the RNA, or perhaps preloading the protein onto the RNA, before its binding site has been transcribed, in preparation for regulation. It would be interesting to try to detect any evidence for this

preloading with a new construct that includes the 5' hairpin structure. Finally, the TRAP protein has been reported have a directional preference in binding (from the 5' to 3' end). Our construct leaves only the 5' end free for binding. If we reverse the sense of the RNA (immobilizing the 5' end or reversing the sequence) can we see a difference in binding? The TRAP protein has been known for more than 20 years, but there are still these important unanswered questions in its function; with single molecule techniques many of these can now, be addressed.

In sum, we have contributed in a small way to all of the necessary components of the design and construction of future self assembling nucleic acid nanomachines. Design, with our forcefield development, experience in the attempt to build a nanostructure, and finally dynamics, with our TRAP studies which may indicate a new method of regulatory interaction. There are still many unanswered questions, but this is the purpose and meaning of scientific inquiry.

Appendix A: Resampling Error Analysis

Resampling is a powerful non-parametric statistical technique[72]. The fundamental insight is to use the data that has been collected to recreate the underlying distribution from which the data was derived. This is possible by repetitively resampling the data from previous experiments to mimic what would be collected if more experiments had been done. If the original data is a fair sample of the parent distribution (*e.g.* the experimental data samples the full range of values in the underlying distribution), reuse of it will allow assessment of how well the experiment has defined the distribution, and what comprises the real physical distribution. The advantage of this methodology is that the resampled error can be found for any function of the data and so analytical propagation of error is not necessary, the resampling is done on the final metric of interest. Another advantage is that the methodology makes no assumptions about the underlying distribution and so performs better in terms of error assessment on distributions that are not Gaussian, and as well as parametric statistics on normal distributions. The drawback is that typically more data is needed for reliable statistical inference as the only information is the data itself (parametric statistics be used on smaller data sets by assuming the distribution is Gaussian), and the computational burden can be very high. These advantages are useful in single molecule work where the amount of data generated is large but the underlying distribution is unknown and propagation of error from measured quantities would be difficult or impossible. There are different ways of implementing the resampling philosophy; the one chosen for this work is the bootstrap method, first introduced by Bradley Efron in 1979 [73]. To conduct this resampling method, one simply makes synthetic data sets from the original data by randomly selecting the

same number of data points with replacement from the original data set. If this is done many times, and the resulting distribution (*i.e.* the distribution created from all of the resampled data sets) will mirror the underlying real distribution (Efron suggests that 25-200 replicates are sufficient for most data sets for statistics on the measure of interest[72]). The standard error of the data set is then estimated by the standard deviation of the resampled sets. If one wants to construct and use confidence intervals for hypothesis testing more sampling is required (usually on the order of 500-1000 samples[72]). This is due to the dependence on the tails of the distribution where there is less sampled data. One suggested measure of when to stop resampling is when the standard error is stable (the coefficient of variation of the standard error reaches a minimum where more sampling makes no improvements).

We implemented a resampling algorithm in IGOR pro to resample our measure of interest, the ratio of the high fret state in the TRAP/tryptophan/RNA titration. In the case of titrations that did not have a ratio, the mean value for the histogram was the measure of interest. The resampling was found to be stable in terms of the standard error with 1000 samples. This was confirmed by comparison of selected points to resampled distributions of 10,000 samples. The error bars present in the graphs are plus and minus the standard error, centered on the point of interest.

IGOR pro Code for Resampling Single Molecule Data

Function resample()

```
    String fileName
    Variable filecount
    String pathName
    String listresamplevalues
    Variable indexm
    Variable FreqTot
    Variable i
    Variable j
    Variable randomnum
    Variable numresamples
    Variable resamplevaluebuffer
    Wave Freq,load0
    numresamples=10000
    Make/D/O/N=(numresamples) resamplevalues

    NewPath/O marcus_path, "G:080610:slide1:resample:"
    indexm=0
            do

                    fileName = IndexedFile(marcus_path, indexm, ".dat")
                    indexm += 1

            while(strlen(fileName) )
    filecount=indexm-1
    print "there are ",filecount,"files"
    indexm = 0

    for(j=0;j<numresamples;j+=1)
            for(i=0;i<filecount;i+=1)
                    randomnum=round(abs(enoise(filecount-1)))
                    // print randomnum
                    fileName = IndexedFile(marcus_path, randomnum, ".dat")
                    // print fileName
                    LoadWave/J/Q/D/N=load/G/K=0/L={0,0,0,1,1}/P=marcus_path
fileName
                    Freq+=load0

            endfor
    SetScale/I x -0.1,1.1,"", Freq
    FreqTot=sum(Freq)
    Freq/=FreqTot
    Halfit(Freq,Freq2)
    resamplevaluebuffer=sum(Freq2, .6,1.1)/(sum(Freq2, -.1,1.1))
    // print resamplevaluebuffer
    resamplevalues[j] = resamplevaluebuffer
```

```
        endfor
End
```

Appendix B: Mathematica Code for the Probabilistic Binding Model

Calculation of the Trp configuration probabilities and corresponding RNA binding "probabilities"(*"probabilities" (= exponentials of minus binding energy divided by the temperature) of the RNA binding to a TRAP site depending on the local Trp configurations (with respect to a configuration without Trps): no Trps, a Trp on a neighboring site, two Trps on two neighboring sites, a Trp on the site, Trps on the site and on the neighboring site, Trps on the site and two neighboring sites*)

v={1,site0neighbor1,site0neighbor2,site1neighbor0,site1neighbor1,site1neighbor 2};nSites=11;stateProb=Table[1,{2^nSites}];RNABindProb=Table[1,{2^nSites}];pRNA BindArray=Table[0,{nSites}];

Do[(*trpconfig is the Trp configuration represented as an array (Table) of length nSites whose elements are 0 or 1 (no Trp and a Trp on the site, respectively)*)trpconfig=IntegerDigits[Trpstate,2,nSites];(**)TrpPairsNr=0;

Do[x=0;(*counting the number of Trp pairs on neighboring sites*)If[trpconfig[[j]]*trpconfig[[Mod[j,nSites]+1]]==1,TrpPairsNr++];

(*"probability" (equilibrium constant) of RNA binding to the site j in the Trp configuration represented by Trpstate or trpconfig*) (*this "probability" depends on the presence of Trps on the sites j, j-1, and j+1; Mod[] is necessary because sites 0 and nSites are equivalent*)

pRNABindArray[[j]]=v[[1+3*trpconfig[[j]]+(trpconfig[[Mod[j-2,nSites]+1]]+trpconfig[[Mod[j,nSites]+1]])]],{j,1,nSites}];

(*total "probability" to bind an RNA to the configuration Trpstate "*)

RNABindProb[[Trpstate+1]]=Product[pRNABindArray[[j]],{j,1,nSites}];

(*probability of the Trp state if pTrp is probability that a site is occupied by the Trp, and pTrpTrp measures Trp-Trp cooperativity*)stateProb[[Trpstate+1]]=pTrp^Total[trpconfig]*(1-pTrp)^(nSites-Total[trpconfig])*pTrpTrp^TrpPairsNr,{Trpstate,0,2^nSites-1}];

Result for the RNA binding probability (stateProb was not normalized!) , a long analytical expression

result=Simplify[Sum[stateProb[[Trpstate+1]]*RNABindProb[[Trpstate+1]],{Trpstate,0,2^nSites-1}]]/Simplify[Total[stateProb]]

((1-pTrp)^11+11 (-1+pTrp)^10 pTrp site0neighbor1^2 site1neighbor0-33 (-1+pTrp)^9 pTrp^2 site0neighbor1^4 site1neighbor0^2-11 (-1+pTrp)^9 pTrp^2 site0neighbor1^2 site0neighbor2 site1neighbor0^2+22 (-1+pTrp)^8 pTrp^3 site0neighbor1^6 site1neighbor0^3+44 (-1+pTrp)^8 pTrp^3 site0neighbor1^4 site0neighbor2 site1neighbor0^3+11 (-1+pTrp)^8 pTrp^3 site0neighbor1^2 site0neighbor2^2 site1neighbor0^3-11 (-1+pTrp)^7 pTrp^4 site0neighbor1^6 site0neighbor2 site1neighbor0^4-33 (-1+pTrp)^7 pTrp^4 site0neighbor1^4 site0neighbor2^2 site1neighbor0^4-11 (-1+pTrp)^7 pTrp^4 site0neighbor1^2 site0neighbor2^3 site1neighbor0^4+11 (-1+pTrp)^6 pTrp^5 site0neighbor1^2 site0neighbor2^4 site1neighbor0^5-11 (-1+pTrp)^9 pTrp^2 pTrpTrp site0neighbor1^2 site1neighbor1^2+55 (-1+pTrp)^8 pTrp^3 pTrpTrp site0neighbor1^4 site1neighbor0 site1neighbor1^2+22 (-1+pTrp)^8 pTrp^3 pTrpTrp site0neighbor1^2 site0neighbor2 site1neighbor0 site1neighbor1^2-33 (-1+pTrp)^7 pTrp^4 pTrpTrp site0neighbor1^6 site1neighbor0^2 site1neighbor1^2-99 (-1+pTrp)^7 pTrp^4 pTrpTrp site0neighbor1^4 site0neighbor2 site1neighbor0^2 site1neighbor1^2-33 (-1+pTrp)^7 pTrp^4 pTrpTrp

site0neighbor1^2 site0neighbor2^2 site1neighbor0^2 site1neighbor1^2+66 (-1+pTrp)^6

pTrp^5 pTrpTrp site0neighbor1^4 site0neighbor2^2 site1neighbor0^3

site1neighbor1^2+44 (-1+pTrp)^6 pTrp^5 pTrpTrp site0neighbor1^2 site0neighbor2^3

site1neighbor0^3 site1neighbor1^2-11 (-1+pTrp)^5 pTrp^6 pTrpTrp site0neighbor2^5

site1neighbor0^4 site1neighbor1^2-22 (-1+pTrp)^7 pTrp^4 pTrpTrp^2 site0neighbor1^4

site1neighbor1^4-11 (-1+pTrp)^7 pTrp^4 pTrpTrp^2 site0neighbor1^2 site0neighbor2

site1neighbor1^4+11 (-1+pTrp)^6 pTrp^5 pTrpTrp^2 site0neighbor1^6 site1neighbor0

site1neighbor1^4+66 (-1+pTrp)^6 pTrp^5 pTrpTrp^2 site0neighbor1^4 site0neighbor2

site1neighbor0 site1neighbor1^4+33 (-1+pTrp)^6 pTrp^5 pTrpTrp^2 site0neighbor1^2

site0neighbor2^2 site1neighbor0 site1neighbor1^4-66 (-1+pTrp)^5 pTrp^6 pTrpTrp^2

site0neighbor1^2 site0neighbor2^3 site1neighbor0^2 site1neighbor1^4-11 (-1+pTrp)^5

pTrp^6 pTrpTrp^3 site0neighbor1^4 site0neighbor2 site1neighbor1^6-11 (-1+pTrp)^5

pTrp^6 pTrpTrp^3 site0neighbor1^2 site0neighbor2^2 site1neighbor1^6+11 (-1+pTrp)^4

pTrp^7 pTrpTrp^3 site0neighbor2^4 site1neighbor0 site1neighbor1^6+11 (-1+pTrp)^8

pTrp^3 pTrpTrp^2 site0neighbor1^2 site1neighbor1^2 site1neighbor2-44 (-1+pTrp)^7

pTrp^4 pTrpTrp^2 site0neighbor1^4 site1neighbor0 site1neighbor1^2 site1neighbor2-22

(-1+pTrp)^7 pTrp^4 pTrpTrp^2 site0neighbor1^2 site0neighbor2 site1neighbor0

site1neighbor1^2 site1neighbor2+11 (-1+pTrp)^6 pTrp^5 pTrpTrp^2 site0neighbor1^6

site1neighbor0^2 site1neighbor1^2 site1neighbor2+66 (-1+pTrp)^6 pTrp^5 pTrpTrp^2

site0neighbor1^4 site0neighbor2 site1neighbor0^2 site1neighbor1^2 site1neighbor2+33

(-1+pTrp)^6 pTrp^5 pTrpTrp^2 site0neighbor1^2 site0neighbor2^2 site1neighbor0^2

site1neighbor1^2 site1neighbor2-44 (-1+pTrp)^5 pTrp^6 pTrpTrp^2 site0neighbor1^2

site0neighbor2^3 site1neighbor0^3 site1neighbor1^2 site1neighbor2+33 (-1+pTrp)^6

pTrp^5 pTrpTrp^3 site0neighbor1^4 site1neighbor1^4 site1neighbor2+22 (-1+pTrp)^6

pTrp^5 pTrpTrp^3 site0neighbor1^2 site0neighbor2 site1neighbor1^4 site1neighbor2-66

(-1+pTrp)^5 pTrp^6 pTrpTrp^3 site0neighbor1^4 site0neighbor2 site1neighbor0

site1neighbor1^4 site1neighbor2-66 (-1+pTrp)^5 pTrp^6 pTrpTrp^3 site0neighbor1^2

site0neighbor2^2 site1neighbor0 site1neighbor1^4 site1neighbor2+33 (-1+pTrp)^4

pTrp^7 pTrpTrp^3 site0neighbor2^4 site1neighbor0^2 site1neighbor1^4

site1neighbor2+33 (-1+pTrp)^4 pTrp^7 pTrpTrp^4 site0neighbor1^2 site0neighbor2^2

site1neighbor1^6 site1neighbor2-11 (-1+pTrp)^7 pTrp^4 pTrpTrp^3 site0neighbor1^2

site1neighbor1^2 site1neighbor2^2+33 (-1+pTrp)^6 pTrp^5 pTrpTrp^3 site0neighbor1^4

site1neighbor0 site1neighbor1^2 site1neighbor2^2+22 (-1+pTrp)^6 pTrp^5 pTrpTrp^3

site0neighbor1^2 site0neighbor2 site1neighbor0 site1neighbor1^2 site1neighbor2^2-33 (-
1+pTrp)^5 pTrp^6 pTrpTrp^3 site0neighbor1^4 site0neighbor2 site1neighbor0^2

site1neighbor1^2 site1neighbor2^2-33 (-1+pTrp)^5 pTrp^6 pTrpTrp^3 site0neighbor1^2

site0neighbor2^2 site1neighbor0^2 site1neighbor1^2 site1neighbor2^2+11 (-1+pTrp)^4

pTrp^7 pTrpTrp^3 site0neighbor2^4 site1neighbor0^3 site1neighbor1^2

site1neighbor2^2-33 (-1+pTrp)^5 pTrp^6 pTrpTrp^4 site0neighbor1^4 site1neighbor1^4

site1neighbor2^2-33 (-1+pTrp)^5 pTrp^6 pTrpTrp^4 site0neighbor1^2 site0neighbor2

site1neighbor1^4 site1neighbor2^2+99 (-1+pTrp)^4 pTrp^7 pTrpTrp^4 site0neighbor1^2

site0neighbor2^2 site1neighbor0 site1neighbor1^4 site1neighbor2^2-22 (-1+pTrp)^3

pTrp^8 pTrpTrp^5 site0neighbor2^3 site1neighbor1^6 site1neighbor2^2+11 (-1+pTrp)^6

pTrp^5 pTrpTrp^4 site0neighbor1^2 site1neighbor1^2 site1neighbor2^3-22 (-1+pTrp)^5

pTrp^6 pTrpTrp^4 site0neighbor1^4 site1neighbor0 site1neighbor1^2 site1neighbor2^3-

22 (-1+pTrp)^5 pTrp^6 pTrpTrp^4 site0neighbor1^2 site0neighbor2 site1neighbor0

site1neighbor1^2 site1neighbor2^3+33 (-1+pTrp)^4 pTrp^7 pTrpTrp^4 site0neighbor1^2 site0neighbor2^2 site1neighbor0^2 site1neighbor1^2 site1neighbor2^3+22 (-1+pTrp)^4 pTrp^7 pTrpTrp^5 site0neighbor1^4 site1neighbor1^4 site1neighbor2^3+44 (-1+pTrp)^4 pTrp^7 pTrpTrp^5 site0neighbor1^2 site0neighbor2 site1neighbor1^4 site1neighbor2^3-44 (-1+pTrp)^3 pTrp^8 pTrpTrp^5 site0neighbor2^3 site1neighbor0 site1neighbor1^4 site1neighbor2^3-11 (-1+pTrp)^5 pTrp^6 pTrpTrp^5 site0neighbor1^2 site1neighbor1^2 site1neighbor2^4+11 (-1+pTrp)^4 pTrp^7 pTrpTrp^5 site0neighbor1^4 site1neighbor0 site1neighbor1^2 site1neighbor2^4+22 (-1+pTrp)^4 pTrp^7 pTrpTrp^5 site0neighbor1^2 site0neighbor2 site1neighbor0 site1neighbor1^2 site1neighbor2^4-11 (-1+pTrp)^3 pTrp^8 pTrpTrp^5 site0neighbor2^3 site1neighbor0^2 site1neighbor1^2 site1neighbor2^4-55 (-1+pTrp)^3 pTrp^8 pTrpTrp^6 site0neighbor1^2 site0neighbor2 site1neighbor1^4 site1neighbor2^4+11 (-1+pTrp)^4 pTrp^7 pTrpTrp^6 site0neighbor1^2 site1neighbor1^2 site1neighbor2^5-22 (-1+pTrp)^3 pTrp^8 pTrpTrp^6 site0neighbor1^2 site0neighbor2 site1neighbor0 site1neighbor1^2 site1neighbor2^5+33 (-1+pTrp)^2 pTrp^9 pTrpTrp^7 site0neighbor2^2 site1neighbor1^4 site1neighbor2^5-11 (-1+pTrp)^3 pTrp^8 pTrpTrp^7 site0neighbor1^2 site1neighbor1^2 site1neighbor2^6+11 (-1+pTrp)^2 pTrp^9 pTrpTrp^7 site0neighbor2^2 site1neighbor0 site1neighbor1^2 site1neighbor2^6+11 (-1+pTrp)^2 pTrp^9 pTrpTrp^8 site0neighbor1^2 site1neighbor1^2 site1neighbor2^7-11 (-1+pTrp) pTrp^10 pTrpTrp^9 site0neighbor2 site1neighbor1^2 site1neighbor2^8+pTrp^11 pTrpTrp^11 site1neighbor2^11)/(1+11 pTrp^2 (-1+pTrpTrp)+11 pTrp^3 (-1+pTrpTrp)^2+11 pTrp^4 (-1+pTrpTrp)^2 (3+pTrpTrp)+11 pTrp^5 (-1+pTrpTrp)^3 (6+pTrpTrp)+11 pTrp^6 (-1+pTrpTrp)^3 (-1+7 pTrpTrp+pTrpTrp^2)+11 pTrp^7 (-1+pTrpTrp)^4 (6+8 pTrpTrp+pTrpTrp^2)+11

pTrp^10 (-1+pTrpTrp)^5 pTrpTrp (-3+3 pTrpTrp^2+pTrpTrp^3)+11 pTrp^9 (-1+pTrpTrp)^5 (-2+5 pTrpTrp+6 pTrpTrp^2+pTrpTrp^3)+11 pTrp^8 (-1+pTrpTrp)^4 (-6+3 pTrpTrp+7 pTrpTrp^2+pTrpTrp^3)+pTrp^11 (-1+pTrpTrp)^6 (-1-6 pTrpTrp+pTrpTrp^2+10 pTrpTrp^3+6 pTrpTrp^4+pTrpTrp^5))

REFERENCES

1. Donchin, Y.; Gopher, D.; Olin, M.; Badihi, Y.; Biesky, M.; Sprung, C. L.; Pizov, R.; Cotev, S., A look into the nature and causes of human errors in the intensive care unit. *Quality and Safety in Health Care* **2003,** *12* (2), 143-147.

2. (a) Michel, F.; Feral, J., Structure and Activities of Group II Introns. *Annual Review of Biochemistry* **1995,** *64* (1), 435-461; (b) Lambowitz, A. M.; Zimmerly, S., Mobile Group II Introns. *Annual Review of Genetics* **2004,** *38* (1), 1-35.

3. Ramakrishnan, V., Histone Structure and the Organization of the Nucleosome. *Annual Review of Biophysics and Biomolecular Structure* **1997,** *26* (1), 83-112.

4. Liu, Q.; Paroo, Z., Biochemical Principles of Small RNA Pathways. *Annual Review of Biochemistry* **2010,** *79* (1), 295-319.

5. (a) Gibson, D. G.; Glass, J. I.; Lartigue, C.; Noskov, V. N.; Chuang, R.-Y.; Algire, M. A.; Benders, G. A.; Montague, M. G.; Ma, L.; Moodie, M. M.; Merryman, C.; Vashee, S.; Krishnakumar, R.; Assad-Garcia, N.; Andrews-Pfannkoch, C.; Denisova, E. A.; Young, L.; Qi, Z.-Q.; Segall-Shapiro, T. H.; Calvey, C. H.; Parmar, P. P.; Hutchison, C. A.; Smith, H. O.; Venter, J. C., Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **2010,** *329* (5987), 52-56; (b) Gibson, D.; Benders, G.; Andrews-Pfannkoch, C.; Denisova, E.; Baden-Tillson, H.; Zaveri, J.; Stockwell, T.; Brownley, A.; Thomas, D.; Algire, M., Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science's STKE* **2008,** *319* (5867), 1215; (c) Gibson, D. G.; Young, L.; Chuang, R.-Y.; Venter, J. C.; Hutchison, C. A.; Smith, H. O., Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Meth* **2009,** *6* (5), 343-345; (d) Lartigue, C.; Vashee, S.; Algire, M. A.;

Chuang, R.-Y.; Benders, G. A.; Ma, L.; Noskov, V. N.; Denisova, E. A.; Gibson, D. G.; Assad-Garcia, N.; Alperovich, N.; Thomas, D. W.; Merryman, C.; Hutchison, C. A.; Smith, H. O.; Venter, J. C.; Glass, J. I., Creating Bacterial Strains from Genomes That Have Been Cloned and Engineered in Yeast. *Science* **2009,** *325* (5948), 1693-1696.

6.   (a) SantaLucia, J., Jr., A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-neighbor Thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1998,** *95*, 1460-1465; (b) SantaLucia, J., Jr.; Hicks, D., The Thermodynamics of DNA Structural Motifs. *Ann. Rev. Biophys. Biomol. Struct.* **2004,** *33*, 413-38; (c) Xia, T.; SantaLucia, J., Jr.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H., Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **1998,** *37*, 14719-14735.

7.   (a) Vriend, G., WHAT IF: a molecular modeling and drug design program. *Journal of Molecular Graphics* **1990,** *8* (1), 52-56; (b) Naito, Y.; Yamada, T.; Ui-Tei, K.; Morishita, S.; Saigo, K., siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference. *Nucleic acids research* **2004,** *32* (suppl 2), W124; (c) Andronescu, M.; Aguirre-Hernandez, R.; Condon, A.; Hoos, H., RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucleic acids research* **2003,** *31* (13), 3416; (d) Ding, Y.; Chan, C.; Lawrence, C., Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic acids research* **2004,** *32* (suppl 2), W135; (e) Zuker, M., Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research* **2003,** *31* (13), 3406-3415.

8.   Shapiro, B. A.; Yingling, Y. G.; Kasprzak, W.; Bindewald, E., Bridging the gap in RNA structure prediction. *Curr. Opin. Struct. Biol.* **2007,** *17* (2), 157-65.

9.   Tinoco, I., Jr.; Bustamante, C., How RNA Folds. *J. Mol. Biol.* **1999,** *293*, 271-281.

10.   Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H., Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *J. Mol. Biol.* **1999,** *288*, 911-940.

11.   Woodson, S. A., Recent insights on RNA folding mechanisms from catalytic RNA. *Cellular and Molecular Life Sciences* **2000,** *57* (5), 796-808.

12.   Sclavi, B.; Sullivan, M.; Chance, M.; Brenowitz, M.; Woodson, S., RNA folding at millisecond intervals by synchrotron hydroxyl radical footprinting. *Science* **1998,** *279* (5358), 1940.

13.   Saenger, W., In Cantor, CR (Ed.) Principles of Nucleic Acid Structure. Springer advanced texts in chemistry. Springer-Verlag, New York, NY: 1984.

14.   Richardson, J. S.; Schneider, B.; Murray, L. W.; Kapral, G. J.; Immormino, R. M.; Headd, J. J.; Richardson, D. C.; Ham, D.; Hershkovits, E.; Williams, L. D.; Keating, K. S.; Pyle, A. M.; Micallef, D.; Westbrook, J.; Berman, H. M., RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* **2008,** *14* (3), 465-481.

15.   (a) Herschlag, D., Biophysical, chemical, and functional probes of RNA structure, interactions and folding: Part A. Preface. *Methods Enzymol.* **2009,** *468*, xv; (b) Draper, D. E., RNA folding: thermodynamic and molecular descriptions of the roles of ions. *Biophys. J.* **2008,** *95*, 5489-5495.

16.	(a) Packer, M.; Hunter, C., Sequence-dependent DNA structure: the role of the sugar-phosphate backbone 1. *Journal of molecular biology* **1998,** *280* (3), 407-420; (b) Schneider, B.; Neidle, S.; Berman, H., Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers* **1997,** *42* (1), 113-124; (c) Sundaralingam, M., Stereochemistry of nucleic acids and their constituents. IV. Allowed and preferred conformations of nucleosides, nucleoside mono-, di-, tri-, tetraphosphates, nucleic acids and polynucleotides. *Biopolymers* **1969,** *7* (6), 821-860; (d) Erfurth, S.; Kiser, E.; Peticolas, W., Determination of the backbone structure of nucleic acids and nucleic acid oligomers by laser Raman scattering. *Proceedings of the National Academy of Sciences* **1972,** *69* (4), 938; (e) Gelbin, A.; Schneider, B.; Clowney, L.; Hsieh, S.; Olson, W.; Berman, H., Geometric parameters in nucleic acids: sugar and phosphate constituents. *J. Am. Chem. Soc* **1996,** *118* (3), 519-529.

17.	Westhof, E.; Fritsch, V., RNA folding: beyond Watson-Crick pairs. *Structure* **2000,** *8* (3), R55-R65.

18.	Quigley, G.; Teeter, M.; Rich, A., Structural analysis of spermine and magnesium ion binding to yeast phenylalanine transfer RNA. *Proceedings of the National Academy of Sciences of the United States of America* **1978,** *75* (1), 64.

19.	(a) Mahen, E.; Harger, J.; Calderon, E.; Fedor, M., Kinetics and thermodynamics make different contributions to RNA folding in vitro and in yeast. *Molecular cell* **2005,** *19* (1), 27-37; (b) Herschlag, D., RNA chaperones and the RNA folding problem. *Journal of Biological Chemistry* **1995,** *270* (36), 20871.

20.   Lu, H.; Skolnick, J., A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins: Structure, Function, and Bioinformatics* **2001,** *44* (3), 223-232.

21.   (a) Treiber, D. K.; Williamson, J. R., Exposing the kinetic traps in RNA folding. *Current Opinion in Structural Biology* **1999,** *9* (3), 339-345; (b) Treiber, D.; Rook, M.; Zarrinkar, P.; Williamson, J., Kinetic intermediates trapped by native interactions in RNA folding. *Science* **1998,** *279* (5358), 1943; (c) Treiber, D.; Williamson, J., Beyond kinetic traps in RNA folding. *Current Opinion in Structural Biology* **2001,** *11* (3), 309-314; (d) Pan, T.; Sosnick, T. R., Intermediates and kinetic traps in the folding of a large ribozyme revealed by circular dichroism and UV absorbance spectroscopies and catalytic activity. *Nat Struct Mol Biol* **1997,** *4* (11), 931-938; (e) Treiber, D. K.; Rook, M. S.; Zarrinkar, P. P.; Williamson, J. R., Kinetic Intermediates Trapped by Native Interactions in RNA Folding. *Science* **1998,** *279* (5358), 1943-1946.

22.   (a) Gelin, B. R.; Karplus, M., Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. *Biochemistry* **1979,** *18* (7), 1256-1268; (b) Vedani, A.; Dunitz, J. D., Lone-pair directionality in hydrogen-bond potential functions for molecular mechanics calculations: the inhibition of human carbonic anhydrase II by sulfonamides. *Journal of the American Chemical Society* **1985,** *107* (25), 7653-7658; (c) Gelin, B. R.; Karplus, M., Sidechain torsional potentials and motion of amino acids in porteins: bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* **1975,** *72* (6), 2002-2006; (d) Das, R.; Baker, D., Macromolecular modeling with rosetta. **2008**.

23.   Buckingham, R. A., The Classical Equation of State of Gaseous Helium, Neon and Argon. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **1938,** *168* (933), 264-283.

24.   Jones, J. E., On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proceedings of the Royal Society of London. Series A* **1924,** *106* (738), 463-477.

25.   Herzfeld, K. F.; Mayer, M. G., On the Theory of Fusion. *Physical Review* **1934,** *46* (11), 995.

26.   (a) Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **2005,** *26* (16), 1668-1688; (b) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995,** *91* (1-3), 1-41.

27.   (a) Peng Yin, H. Y., Xiaoju G. Daniell, Andrew J. Turberfield, John H. Reif, A Unidirectional DNA Walker That Moves Autonomously along a Track. *Angewandte Chemie International Edition* **2004,** *43*, 4906 –4911; (b) Seeman, W. B. S. a. N. C., A Precisely Controlled DNA Biped Walking Device. *Nano Letters* **2004,** *4* (7), 1203-1207; (c) Liu, X. L. a. D. R., DNA-Templated Organic Synthesis: Nature's Strategy for Controlling Chemical Reactivity Applied to Synthetic Molecules. *Angewandte Chemie International Edition* **2004,** *43*, 4848 – 4870.

28.   (a) Fujibayashi, K.; Hariadi, R.; Park, S. H.; Winfree, E.; Murata, S., Toward Reliable Algorithmic Self-Assembly of DNA Tiles: A Fixed-Width Cellular Automaton Pattern. *Nano Lett.* **2007**; (b) Goodman, R. P.; Schaap, I. A. T.; Tardin, C. F.; Erben, C. M.; Berry, R. M.; Schmidt, C. F.; Turberfield, A. J., Rapid Chiral Assembly of Rigid DNA Building Blocks for Molecular Nanofabrication. *Science* **2005,** *310* (5754), 1661-1665.

29.   J.Rose, L. M. W. a. S., A novel DNA joining activity catalyzed by T4 DNA ligase. *Nucleic Acids Research* **1991,** *19* (4), 809-813.

30.   Barany, F., Genetic disease detection and DNA amplification using cloned thermostable ligase. *Proceeding of the National Academy of Sciences of the United States of America* **1991,** *88*, 189-193.

31.   Hsueh, C.; Chen, H.; Gimzewski, J. K.; Reed, J.; Abdel-Fattah, T. M., Localized Nanoscopic Surface Measurements of Nickel-Modified Mica for Single-Molecule DNA Sequence Sampling. *ACS Applied Materials & Interfaces* **2010,** *2* (11), 3249-3256.

32.   Hopkins, P.; Millard, J.; Woo, J.; Weidner, M.; Kirchner, J.; Sigurdsson, S.; Raucher, S., Sequence preferences of DNA interstrand cross-linking agents: importance of minimal DNA structural reorganization in the cross-linking reactions of mechlorethamine, cisplatin and mitomycin C. *Tetrahedron* **1991,** *47* (14-15), 2475-2489.

33.   Anderson, T. F.; Voorhees, J. J., Psoralen Photochemotherapy of Cutaneous Disorders. *Annual Review of Pharmacology and Toxicology* **1980,** *20* (1), 235-257.

34.   Cimino, G.; Gamper, H.; Isaacs, S.; Hearst, J., Psoralens as photoactive probes of nucleic acid structure and function: organic chemistry, photochemistry, and biochemistry. *Annual review of biochemistry* **1985,** *54* (1), 1151-1193.

35.   Daniell, M. D.; Hill, J. S., A HISTORY OF PHOTODYNAMIC THERAPY. *Australian and New Zealand Journal of Surgery* **1991,** *61* (5), 340-348.

36.   (a) Mattes, W.; Hartley, J.; Kohn, K., DNA sequence selectivity of guanine-N7 alkylation by nitrogen mustards. *Nucleic acids research* **1986,** *14* (7), 2971; (b) Kohn, K.; Hartley, J.; Mattes, W., Mechanisms of DNA sequence selective alkylation of guanine-N7 positions by nitrogen mustards. *Nucleic acids research* **1987,** *15* (24), 10531.

37.   George D. Cimino, H. B. G., Stephen T. Isaacs, John E. Hearst, PSORALENS AS PHOTOACTIVE PROBES OF NUCLEIC ACID STRUCTURE AND FUNCTION: ORGANIC CHEMISTRY, PHOTOCHEMISTRY, AND BIOCHEMISTRY. *Annual Review of Biochemistry* **1985,** *54*, 1151-1193.

38.   George D. Cimino, Y.-b. S., and John E. Hearst, Wavelength Dependence for the Photoreversal of a Psoralen-DNA Cross-Link. *Biochemistry* **1986,** *25*, 3013-3020.

39.   Berkenkamp, S.; Kirpekar, F.; Hillenkamp, F., Infrared MALDI mass spectrometry of large nucleic acids. *Science* **1998,** *281* (5374), 260.

40.   Bensimon, D.; Simon, A.; Croquette, V.; Bensimon, A., Stretching DNA with a receding meniscus: experiments and models. *Physical review letters* **1995,** *74* (23), 4754-4757.

41.   Jiang, Y.; Rabbi, M.; Mieczkowski, P. A.; Marszalek, P. E., Separating DNA with Different Topologies by Atomic Force Microscopy in Comparison with Gel Electrophoresis. *The Journal of Physical Chemistry B* **2010,** *114* (37), 12162-12165.

42.   Hansma, H.; Revenko, I.; Kim, K.; Laney, D., Atomic force microscopy of long and short double-stranded, single-stranded and triple-stranded nucleic acids. *Nucleic acids research* **1996,** *24* (4), 713.

43.   Schaak, J. E.; Babitzke, P.; Bevilacqua, P. C., Phylogenetic conservation of RNA secondary and tertiary structure in the trpEDCFBA operon leader transcript in Bacillus. *RNA* **2003,** *9* (12), 1502-1515.

44.   (a) Merino, E.; Babitzke, P.; Yanofsky, C., trp RNA-binding attenuation protein (TRAP)-trp leader RNA interactions mediate translational as well as transcriptional regulation of the Bacillus subtilis trp operon. *Journal of Bacteriology* **1995,** *177* (22), 6362; (b) Yakhnin, H.; Yakhnin, A. V.; Babitzke, P., Translation Control of trpG from Transcripts Originating from the Folate Operon Promoter of Bacillus subtilis Is Influenced by Translation-Mediated Displacement of Bound TRAP, While Translation Control of Transcripts Originating from a Newly Identified trpG Promoter Is Not. *J. Bacteriol.* **2007,** *189* (3), 872-879; (c) Du, H.; Babitzke, P., trp RNA-binding Attenuation Protein-mediated Long Distance RNA Refolding Regulates Translation of trpE in Bacillus subtilis. *Journal of Biological Chemistry* **1998,** *273* (32), 20494; (d) Snyder, D.; Lary, J.; Chen, Y.; Gollnick, P.; Cole, J. L., Interaction of the trp RNA-binding Attenuation Protein (TRAP) with Anti-TRAP. *Journal of Molecular Biology* **2004,** *338* (4), 669-682; (e) Valbuzzi, A.; Yanofsky, C., Inhibition of the B. subtilis Regulatory Protein TRAP by the TRAP-Inhibitory Protein, AT. *Science* **2001,** *293* (5537), 2057-2059.

45.   Yakhnin, H.; Zhang, H.; Yakhnin, A. V.; Babitzke, P., The trp RNA-Binding Attenuation Protein of Bacillus subtilis Regulates Translation of the Tryptophan Transport Gene trpP (yhaG) by Blocking Ribosome Binding. *J. Bacteriol.* **2004,** *186* (2), 278-286.

46.	(a) Valbuzzi, A.; Gollnick, P.; Babitzke, P.; Yanofsky, C., The Anti-trp RNA-binding Attenuation Protein (Anti-TRAP), AT, Recognizes the Tryptophan-activated RNA Binding Domain of the TRAP Regulatory Protein. *Journal of Biological Chemistry* **2002,** *277* (12), 10608-10613; (b) Chen, G.; Yanofsky, C., Tandem Transcription and Translation Regulatory Sensing of Uncharged Tryptophan tRNA. *Science* **2003,** *301* (5630), 211-213; (c) Sarsero, J. P.; Merino, E.; Yanofsky, C., A Bacillus subtilis operon containing genes of unknown function senses tRNATrp charging and regulates expression of the genes of tryptophan biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* **2000,** *97* (6), 2656-2661; (d) Henkin, T. M., Transcription termination control in bacteria. *Current opinion in microbiology* **2000,** *3* (2), 149-153.

47.	Yakhnin, H.; Yakhnin, A. V.; Babitzke, P., The trp RNA-binding attenuation protein (TRAP) of Bacillus subtilis regulates translation initiation of ycbK, a gene encoding a putative efflux protein, by blocking ribosome binding. *Molecular Microbiology* **2006,** *61* (5), 1252-1266.

48.	Gollnick, P.; Babitzke, P.; Antson, A.; Yanofsky, C., COMPLEXITY IN REGULATION OF TRYPTOPHAN BIOSYNTHESIS IN BACILLUS SUBTILIS. *Annual Review of Genetics* **2005,** *39* (1), 47-68.

49.	Antson, A.; Dodson, E.; Dodson, G.; Greaves, R.; Chen, X.; Gollnick, P., Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature* **1999,** *401* (6750), 235-242.

50.	Antson, A. A.; Brzozowski, A. M.; Dodson, E. J.; Dauter, Z.; Wilson, K. S.; Kurecki, T.; Otridge, J.; Gollnick, P., 11-fold Symmetry of the trp RNA-binding

Attenuation Protein (TRAP) from Bacillus subtilis Determined by X-ray Analysis. *Journal of Molecular Biology* **1994,** *244* (1), 1-5.

51.   Babitzke, P.; Bear, D. G.; Yanofsky, C., TRAP, the trp RNA-binding attenuation protein of Bacillus subtilis, is a toroid-shaped molecule that binds transcripts containing GAG or UAG repeats separated by two nucleotides. *Proceedings of the National Academy of Sciences of the United States of America* **1995,** *92* (17), 7916-7920.

52.   McElroy, C.; Manfredo, A.; Gollnick, P.; Foster, M., Thermodynamics of Tryptophan-Mediated Activation of the trp RNA-Binding Attenuation Protein†. *Biochemistry* **2006,** *45* (25), 7844-7853.

53.   (a) Antson, A.; Otridge, J.; Brzozowski, A.; Dodson, E.; Dodson, G.; Wilson, K.; Smith, T.; Yang, M.; Kurecki, T.; Gollnick, P., The structure of trp RNA-binding attenuation protein. **1995**; (b) Yakhnin, A.; Trimble, J.; Chiaro, C.; Babitzke, P., Effects of mutations in the L-tryptophan binding pocket of the trp RNA-binding attenuation protein of Bacillus subtilis. *Journal of Biological Chemistry* **2000,** *275* (6), 4519; (c) Li, P. T. X.; Scott, D. J.; Gollnick, P., Creating Hetero-11-mers Composed of Wild-type and Mutant Subunits to Study RNA Binding to TRAP. *Journal of Biological Chemistry* **2002,** *277* (14), 11838-11844; (d) Li, P. T. X.; Gollnick, P., Characterization of a trp RNA-binding Attenuation Protein (TRAP) Mutant with Tryptophan Independent RNA Binding Activity. *Journal of Molecular Biology* **2004,** *335* (3), 707-722.

54.   Elliott, M.; Gottlieb, P.; Gollnick, P., The mechanism of RNA binding to TRAP: initiation and cooperative interactions. *RNA* **2001,** *7* (1), 85.

55.   McElroy, C.; Manfredo, A.; Wendt, A.; Gollnick, P.; Foster, M., TROSY-NMR Studies of the 91 kDa TRAP Protein Reveal Allosteric Control of a Gene Regulatory

Protein by Ligand-altered Flexibility. *Journal of Molecular Biology* **2002,** *323* (3), 463-473.

56.   Alemán, E.; Lamichhane, R.; Rueda, D., Exploring RNA folding one molecule at a time. *Current Opinion in Chemical Biology* **2008,** *12* (6), 647-654.

57.   Inc., W., *IGOR Pro Version 5*. WaveMetrics Inc.: Lake Oswego, OR, 2005.

58.   Pyle, A., Metal ions in the structure and function of RNA. *Journal of biological inorganic chemistry* **2002,** *7* (7), 679-690.

59.   Aitken, C. E.; Marshall, R. A.; Puglisi, J. D., An Oxygen Scavenging System for Improvement of Dye Stability in Single-Molecule Fluorescence Experiments. *Biophysical Journal* **2008,** *94* (5), 1826-1835.

60.   Rasnik, I.; McKinney, S. A.; Ha, T., Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat Meth* **2006,** *3* (11), 891-893.

61.   Baumann, C.; Otridge, J.; Gollnick, P., Kinetic and Thermodynamic Analysis of the Interaction between TRAP (trp RNA-binding Attenuation Protein) of Bacillus subtilis and trp Leader RNA. *Journal of Biological Chemistry* **1996,** *271* (21), 12269-12274.

62.   McCammon, M. G.; Hernández, H.; Sobott, F.; Robinson, C. V., Tandem Mass Spectrometry Defines the Stoichiometry and Quaternary Structural Arrangement of Tryptophan Molecules in the Multiprotein Complex TRAP. *Journal of the American Chemical Society* **2004,** *126* (19), 5950-5951.

63.   McCabe, B.; Gollnick, P., Cellular levels of trp RNA-binding attenuation protein in Bacillus subtilis. *Journal of Bacteriology* **2004,** *186* (15), 5157.

64.   (a) Babitzke, P.; Yanofsky, C., Reconstitution of Bacillus subtilis trp attenuation in vitro with TRAP, the trp RNA-binding attenuation protein. *Proceedings of the*

National Academy of Sciences of the United States of America **1993,** *90* (1), 133; (b)

Otridge, J.; Gollnick, P., MtrB from Bacillus subtilis binds specifically to trp leader RNA

in a tryptophan-dependent manner. *Proceedings of the National Academy of Sciences*

1993, *90* (1), 128.

65.   Scatchard, G., The attractions of proteins for small molecules and ions. *Annals of*

the New York Academy of Sciences **1949,** *51* (Molecular Interaction), 660-672.

66.   (a) Epstein, I., Cooperative and non-cooperative binding of large ligands to a

finite one-dimensional lattice:: A model for ligand-ougonucleotide interactions.

Biophysical Chemistry **1978,** *8* (4), 327-339; (b) Bujalowski, W.; Lohman, T.; Anderson,

C., On the cooperative binding of large ligands to a one-dimensional homogeneous

lattice: the generalized three-state lattice model. *Biopolymers* **1989,** *28* (9), 1637-1643;

(c) McGhee, J.; von Hippel, P., Theoretical aspects of DNA-protein interactions: Co-

operative and non-co-operative binding of large ligands to a one-dimensional

homogeneous lattice* 1. *Journal of Molecular Biology* **1974,** *86* (2), 469-489.

67.   Wolfram Research, I., *Mathematica Edition: Version 7.0.* Wolfram Research,

Inc.: Champaign, Illinois, 2008.

68.   Yang, M.; Chen, X.-p.; Militello, K.; Hoffman, R.; Fernandez, B.; Baumann, C.;

Gollnick, P., Alanine-scanning mutagenesis of Bacillus subtilis trp RNA-binding

attenuation protein (TRAP) reveals residues involved in tryptophan binding and RNA

binding. *Journal of Molecular Biology* **1997,** *270* (5), 696-710.

69.   (a) Polach, K.; Widom, J., A model for the cooperative binding of eukaryotic

regulatory proteins to nucleosomal target sites. *Journal of Molecular Biology* **1996,** *258*

(5), 800-812; (b) Polach, K.; Widom, J., Mechanism of protein access to specific DNA

sequences in chromatin: a dynamic equilibrium model for gene regulation. *Journal of Molecular Biology* **1995,** *254* (2), 130-149.

70.   (a) Riechmann, L.; Winter, G., Early Protein Evolution: Building Domains from Ligand-binding Polypeptide Segments. *Journal of Molecular Biology* **2006,** *363* (2), 460-468; (b) Riechmann, L.; Winter, G., Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proceedings of the National Academy of Sciences of the United States of America* **2000,** *97* (18), 10068-10073; (c) Riechmann, L.; Lavenir, I.; de Bono, S.; Winter, G., Folding and Stability of a Primitive Protein. *Journal of Molecular Biology* **2005,** *348* (5), 1261-1272; (d) de Bono, S.; Riechmann, L.; Girard, E.; Williams, R. L.; Winter, G., A segment of cold shock protein directs the folding of a combinatorial protein. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (5), 1396-1401; (e) Park, C.; Raines, R., Dimer formation by a "monomeric" protein. *Protein Science* **2000,** *9* (10), 2026-2033.

71.   Davidson, A.; Lumb, K.; Sauer, R., Cooperatively folded proteins in random sequence libraries. *Nature Structural & Molecular Biology* **1995,** *2* (10), 856-864.

72.   Efron, B.; Tibshirani, R.; Tibshirani, R. J., *An introduction to the bootstrap*. Chapman & Hall/CRC: 1993.

73.   Efron, B., Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **1979,** *7* (1), 1-26.

ABSTRACT

ARTIFICIAL AND NATURAL NUCLEIC ACID SELF ASSEMBLING SYSTEMS

by

MARCUS WOOD

August 2011

Advisor: Dr. John SantaLucia, Jr., Dr. David Rueda

Major: Chemistry

Degree: Doctor of Philosophy

Nucleic acids are good candidates for nanomachine construction. They participate in all the processes of life, and so can function as structural building blocks and dynamic catalysts. However, to use nucleic acids as nanomachines, a better understanding of their material properties, how to design structures using them, and their dynamics is needed. We have tried to address these issues, in a small way, with nucleic acid force field development, an attempt at nanostructural design and synthesis using DNA, and a study of the RNA/protein regulatory dynamics of the tryptophan regulatory attenuation protein.

AUTOBIGRAPHICAL STATEMENT

Marcus Wood

I was born in Mexico City, Mexico. As my father's employment involved extensive travel I have lived in many different countries. The condensed version of my peregrinations is Mexico, U.S.A, U.K., U.S.A., New Zealand, and back to the U.S.A. My parents continued to visit quite a few other countries, but I stayed in the U.S. to do my bachelor's at the University of Michigan. After this I worked for two years at the Ann Arbor water treatment plant in operations. I've had many jobs over the years, but I am most proud of my time at the water treatment plant. The job gave me the satisfaction of public service, enough funds to go back to school, and I got to work with a spectacular group of dedicated people. I left the water treatment plant to continue my education with a Master's degree (which was my original intention until my advisor convinced me to enter the Ph.D. program). When I first went to university, I was interested in both physics and chemistry for a degree; I chose chemistry, but it has been satisfying to eventually combine these two interests in my Ph.D. work in biophysics.

Education:

- **Ph.D. in Chemistry (2011)**
  Wayne State University, Detroit, Michigan

- **2001 B.S. in Chemistry (2001)**
  The University of Michigan, Ann Arbor, Michigan