Evolving The Deoxyribozyme-Based Logic Gate Design Process Through MAYA-II Reconstruction

M. Leigh Fanning[1], Joanne Macdonald[2], Darko Stefanovic[1]

[1]Department of Computer Science, University of New Mexico,
Albuquerque, New Mexico 87131

[2]Division of Experimental Therapeutics, Department of Medicine, Columbia University,
New York, New York 10032

May 4, 2009

Abstract

We previously described a tic-tac-toe playing molecular automaton, MAYA-II, constructed from a molecular array of deoxyribozyme-based logic gates, that uses oligonucleotides as inputs and outputs. We are now developing an ensemble modeling tool for high-throughput oligonucleotide input and logic gate designs. The modeling tool is based on exhaustive reconstruction of both intended and unintended reactions between MAYA-II gates and inputs, and seeks to directly correlate empirical observations with computational predictions. Here we describe exhaustive analysis of the MAYA-II *Yes* logic gates folding structures, both alone and in conjunction with the MAYA-II oligonucleotide inputs. Results indicate that *in silico* modeling accurately reflects experimental results, creating a predictive value and benchmark for future high-throughput oligonucleotide input and *Yes* gate designs. These studies serve purpose towards our goal of constructing a generalized oligonucleotide design library for expansion of molecular computation beyond hundreds, to millions of potential interactions, conferring greater functionality in terms of both reliability and complexity.

1. Introduction

Computation on a molecular substrate has been physically realized in automata devised by Adelman [1], Benenson et al. [3], and in our laboratory as the MAYA-I [14] and MAYA-II [17] experiments, among others. MAYA, a molecular array of *Yes* and *And* gates, consists of individual deoxyribozyme-based logic gates and sequentially introduced DNA oligonucleotide inputs. These automata effect control by careful biochemical arrangement. With the ensuing heterogenous population of tens of distinct oligonucleotide species together in solution, successful computation depends on achieving satisfactory thresholds of hundreds of desired events, including gates folding to their intended secondary structure and inputs binding to gates at precisely the correct locations along their entire length. These potential events are positive in the sense that they are necessary to carry out the intended computation, yet commensurately

there are thousands of negative events that must not occur, or occur at low enough rates so as to minimize noise or system failure.

Design efforts hence walk a tight line. In practice, *in silico* planning [2,5,11,12,16,18] has centered around sequence selection that will achieve a particular target secondary structure, matching gate recognition regions to the reverse complements of input sequences, and ensuring sufficient content difference between all other element combinations that may be present in solution but intended to have agnostic relationships. Multiple gates and inputs are then combinable in solution and may carry out parallel hybridization reactions as long as sufficient primary structure differences exist between all gate/gate, input/input, and gate/input tuples designed to not hybridize. We assume both negative and positive affinity relationships. By employing Hamming distance metrics, we assume suitable separation will minimize or eliminate the problem of cross-talk. Correspondingly, we assume that matching gate recognition regions to inputs ensures reliable, measurable oligonucleotide cleavage or hybridization outputs, which are then de facto communication signals carrying information.

A reasonable hypothesis is that these assumptions should be equally borne out in stand-alone DNA models advertising predictiveness and further that empirical results should additionally confirm model predictions. However, as far as we are aware, the empirical feedback from the significant *in vivo* tuning, testing, and modifications required to achieve computational logic goals within a chemical network has not to date been fully incorporated back into the design process.

To this end, we are undertaking an analytical reconstruction of the MAYA-II automaton. MAYA-II now serves as a rich source of unmined data which can be used as design feedback. Reconstruction is a reverse engineering effort, and has clear goals. First, we would like to compare and correlate empirical observations to offline model predictions. Second, we want to be able to deduce sources of noise resulting from unintended reactions and explain varying signal intensity from some of the deoxyribozyme gates. Finally, and most importantly, MAYA-II reconstruction results are useful in developing an ensemble model for future designs and deoxyribozyme library construction. The full details of MAYA-II, including all experimental results, are in [14]. We report here our reconstruction approach, development of our general ensemble modeling tool Pyxis, and reconstruction results in progress which have confirmed some initial expectations. Pyxis exhaustively enumerates all reactions which must be accounted for in a molecular computing system, and allows testing against a suite of existing domain-specific models. Collating ensemble model prediction results, along with empirical data comparison, will permit design scaling from thousands to millions of competing reactions.

1.1 Overview of the MAYA II automaton

Tic-tac-toe is a deterministic game of perfect information, amenable to representation of possible games in a game tree wherein nodes depict orderings of move decisions for each player. The automaton adversary was directed to play all possible games with a perfect non-losing strategy and was given the upper hand by always playing the first move into the center square. All root-to-leaf paths through the game tree ended in either victory or a draw for the automaton. These paths, encoded into a set of Boolean logic formulas, were physically realized in solution as deoxyribozyme-based logic gates.

In total, 76 games were divided into four subsets of 19 games such that each subset represented game tree paths arising from different corner and side plays executed as the second overall move of the game, and first play for the human opponent. Human moves corresponded to introduction of DNA oligonucleotide inputs into a test tube well representing a square of the game. Eight solution wells, representing the eight open squares at game outset, were preloaded with deoxyribozyme-based gate constructs designed to carry out the specific Boolean formulae for human play echo and automaton response. With each input introduction into the system, exactly two gate/input interactions were designed to occur: one in the well where the human intended to mark a square, and the other in a different well where the automaton intended to mark its response. Move detection and display was accomplished via secondary

oligonucleotide substrate cleavage by the gate producing red tetramethylrhodamine (TAMRA) or green fluorescein visible fluorescence. The Δ*F/min* signal directly reflected achieving a threshold concentration of cleavage product, which depended on correct hybridization between input and gate oligonucleotides.

1.2 Handling combinatorics of MAYA-II oligonucleotide populations

A central goal in reconstruction has been to account for all potential interactions that could occur incidentally or deliberately over the course of playing out each game. The notion of interaction involves an ordering of certain events. For each event, we further attempt to estimate levels of participation amongst the different nucleic acid reactants, as well as to quantify products. To get from introduction of an input into each of the wells, to detection of fluorescent output, 1) deoxyribozyme gates must correctly fold to the designed secondary structure stem-loop conformations, 2) inputs must seek out and bind to the gates in the loop regions connected to stem sections, 3) the stem loop must undergo a conformational change wherein it opens, and 4) the exposed remainder of the stem section must bind and cleave substrate molecules. The critical steps are gates attaining expected secondary structure and inputs correctly binding to gates, each of which is designed on the basis of minimization of predicted free energies of the gates and gate/input complex formations. Other relevant factors, which are no less important, are gate and input concentrations, diffusion, temperature, k_{on} and k_{off} binding and unbinding efficiencies, and the chemical kinetics involved for either hybridization or cleavage reactions. These are beyond the scope of the present paper, and will be addressed in the ongoing evolution of our ensemble model.

We have initially focused our efforts on accounting for all the possible ways gates and inputs might come together, as either positive in terms of one or more inputs correctly binding to a gate which has folded properly, or negative for any other combination, including inputs binding to other inputs, gates binding to other gates, or inputs binding to the wrong gates. Figure 1 illustrates a single positive reaction (bold text) competing with a number of negative reactions (plain text).

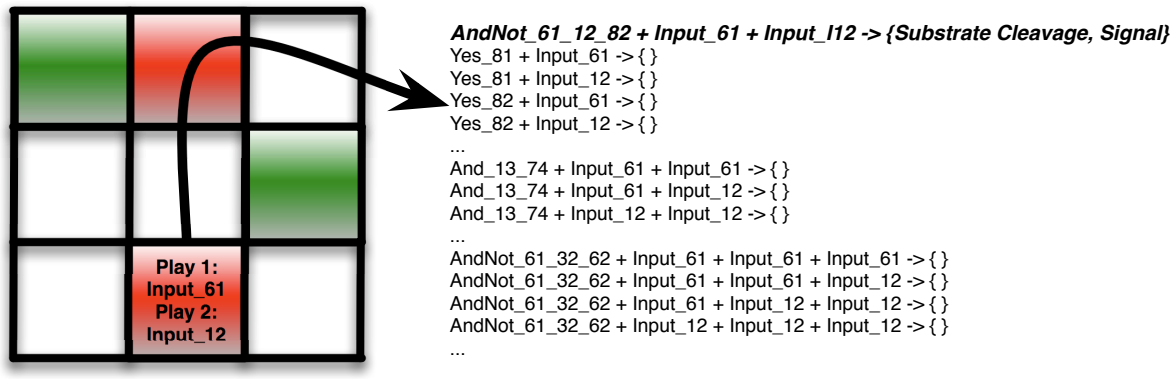

Figure 1: Sampling of possible hybridizations following input introduction. Well 8 Yes gates (one input), And gates (two inputs), and AndNot gates (two inputs, absence of third) are designed to activate or remain quiescent based on their encoded target recognition regions. Direct homology match to inputs I61 and I12 initiates signaling from gate

For a full accounting and individual evaluation, we enumerate potential reactions as follows. Given the sets of all games, all plays within each game, and all active wells, the sets of inputs *I* and gates *G* are:

$$\texttt{wells} = \{1,2,3,4,6,7,8,9\},\ \texttt{games} = \{1,...,76\},\ \texttt{plays} = \{1,2,3,4\},$$

$$I = \{input_{i,j,k} | i \in \texttt{wells}, j \in \texttt{games}, k \in \texttt{plays}\}$$

$$G_{YES} = \{gate_{i,j,k_1} | i \in \texttt{wells}, j \in \texttt{games}, k_1 \in \texttt{plays}\}$$
$$G_{AND} = \{gate_{i,j,k_1,k_2} | i \in \texttt{wells}, j \in \texttt{games}, k_1, k_2 \in \texttt{plays}\}$$
$$G_{ANDANDNOT} = \{gate_{i,j,k_1,k_2,k_3} | i \in \texttt{wells}, j \in \texttt{games}, k_1, k_2, k_3 \in \texttt{plays}\}$$

$$G = G_{YES} \cup G_{AND} \cup G_{ANDANDNOT}$$

where *Yes* Gates admit one input, *And* Gates admit two inputs, and *And-And-Not* Gates admit three inputs. We consider reaction sets I, the set of inputs alone, G, the set of gates alone, $G \times G$, the set of pairs of gates, $G_{YES} \times I$, the set of pairs of *Yes* Gates and inputs, $G_{AND} \times I \times I$, the Cartesian product set of *And* Gates and input pairs, and finally, $G_{ANDANDNOT} \times I \times I \times I$, the Cartesian product set of *And-And-Not* Gates and input triples. For each game, inputs are added to all of the wells, but we index the set of inputs with the well number reflecting where either player desired to mark a well in a particular play.

1.3 Development of Pyxis as an ensemble modeling tool

For evaluating and testing reconstruction combinations, and relating back to empirical evidence, we determined that conceiving of any new DNA domain-specific model or physical simulation was not warranted. Instead, we needed a way to semi-automatically account for the prohibitory complexity arising from using different deoxyribozymes and oligonucleotides together in non-trivial ways. Further we both wished for a more general way to evaluate arbitrary combinations using existing models, and observed that greater functionality chemical computing is moving towards more complicated circuits and more circuits networked together. This implies substantially greater numbers of distinct species, which would prove untenable for exhaustive laboratory testing during development. Additionally, inferring function and behavior from nucleic acid structure alone falls short for all but the very simplest of combinations involving only a few participating elements. System parameters are interdependent and must be addressed as such as closely as possible prior to testing in the laboratory.

Our solution is development of Pyxis, an ensemble modeling tool which serves as a framework for enumerating combinations of elements, testing against a suite of models, collating results and correlating results from different models where possible. Pyxis is encoded as a pure Python application [13] fronting a PostgreSQL database which houses domain model test results and all deoxyribozyme and oligonucleotide elements. Existing models may be dropped in, input interfacing via command line or files is automatically handled, as is harvesting of data from output files. Users are not required to manage the fine details for each model and are directed to provide only the essential details, such as selection of deoxyribozymes from the database, temperature, and similar relevant parameters. Pyxis brokers all model executions and stores results back into the database for later uses such as plotting, computing statistics, and results comparison where appropriate.

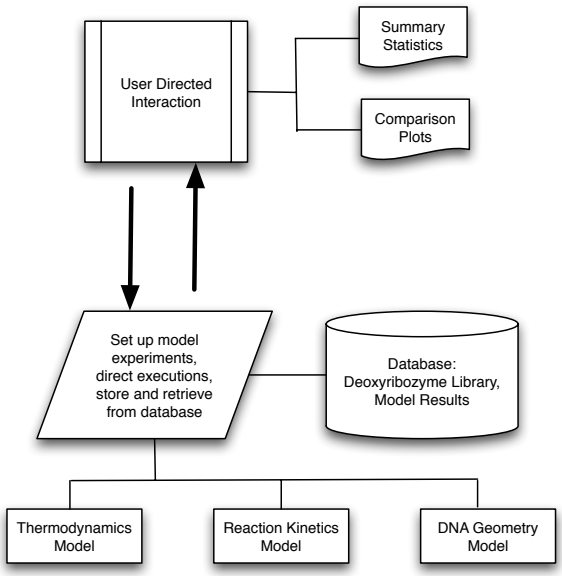

Figure 2: Pyxis architecture. Single or batch experiments are directed towards requested models. Models remain in natively encoded languages, input and output is packaged by the program.

The first model incorporated into Pyxis is NUPACK [6]. NUPACK builds on MFold [18] nucleic acid folding and hybridization prediction software and extends prediction from a single nucleic strand to an arbitrary number of nucleic acid strands. NUPACK was not available during engineering of MAYA-II and is appropriate for understanding the large number of possible hybridizations which we would like to account for. NUPACK develops a multistranded secondary structure model where different strands are placed end to end and evaluated on a circular polymer graph. Permutations of all strand orderings are considered, and for any one ordering dynamic programming recursions are employed to correct for overcounting of free energy contributions from base pairing evaluation.

2 Results

We show results for MAYA-II *Yes* gates generated using Pyxis to both deduce all combinations of gates and inputs, and test each using NUPACK. Two different deoxyribozyme designs were used in the design of the *Yes* Gates. The stem-loop portions of each are displayed in Figure 3. For human moves, the 8.17.1 core was employed which uses a 10 nucleotide (nt) stem around the 15 nt input target subsequence loop. For automaton moves, the E6 core was employed which uses a 6-8 nt stem around the 15 nt input target subsequence loop. The stem region, alternatively termed critical structure, must pair-bind as shown, such that no gaps, puckers or otherwise ill-formed bindings occur, leaving the loop exposed for the later input recognition reaction. The loop region must itself be free of any pair bindings. Attaining these configurations constitutes optimal structure folding.

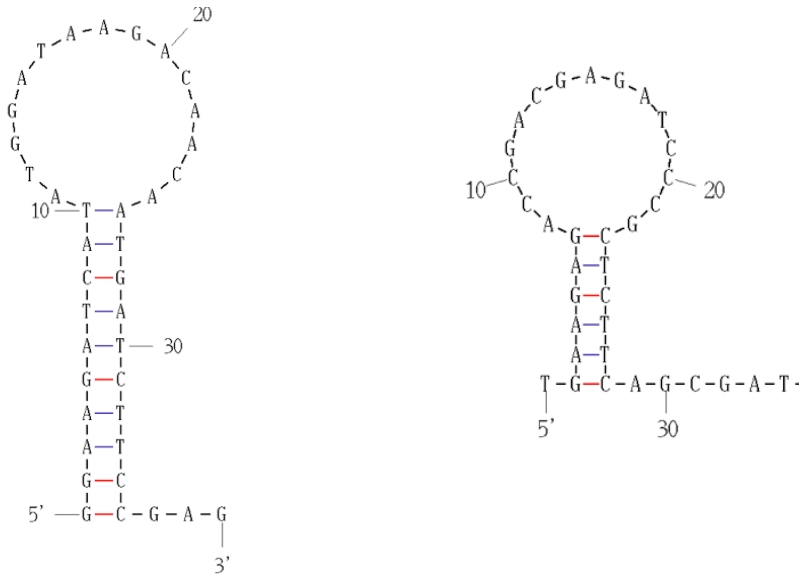

Figure 3: Stem-loop portions of Yes-Gate deoxyribozymes. The 8.17.1 design (left) 10 nt stem supports a 15 nt loop input recognition region. The E-6 design (right) 6 nt stem also supports a 15 nt loop input recognition region. In some cases, input dependent adjustments of the E-6 resulted in 7 and 8 nt stems lengths. Not shown past nt 30 is the core region used for substrate cleavage. Diagrams courtesy of the DinaMelt web based server [15].

All the *Yes* gates were analyzed individually for self-hybridization and optimal folding. The evaluation shows all gates folding properly within the critical structure, over half (55%, Table 1) show small amounts of additional binding within the loop region intended for input recognition. These additional bindings use between two and four pairs of nucleotides within the loop region, and were observed for both core types. There was no preferential location along the loop region for these extra bindings.

Table 1: Number of nucleotide pairs bound within input loop regions

0 pairs	1 pair	2 pairs	3 pairs	4 pairs
18	0	14	5	3

Next, we used Pyxis to run NUPACK and test for cross-hybridization of gates with both intended and unintended inputs. Gates tested with intended inputs constitute a positive set, whereas gates tested with unintended inputs they were not designed for constitute a negative set. For MAYA-II, 40 Yes gates combinations were positive, and 1240 were negative. Despite some gates alone showing a limited tendency to bind within their input regions, when tested together with the intended inputs, the bindings occur perfectly in all cases. The predicted minimum free energy structures MFE for all combinations are summarized in Table 2, and displayed in Figure 4. The difference in average MFE in gates with designed inputs (-28.3 kcal/mol 8.17.1, -31.48 kcal/mol E6) compared to MFE of the gate alone (- 12.13 kcal/mol 8.17.1, -12.45 kcal/mol E6) is significant. We expect this value could be used as a predictive value for indicating activation of a gate, whether through intended or unintended input binding.

Table 2: Minimum Free Energy Secondary Structures

(kcal/mol)	8.17.1 core alone	8.17.1 core with design input	8.17.1 core with incidental input	E6 core alone	E6 core with design input	E6 core with incidental input
Average MFE	-12.13	-28.3	-18.99	-12.45	-31.48	-19.71
St. Dev. MFE	0.67	1.76	1.47	0.81	2.12	1.96
Min MFE	-13.74	-32.76	-23.63	-14.54	-28.23	-26.3
Max MFE	-11.09	-24.23	-15.98	-11.88	-34.83	-15.73

E6 cores combined with unintended inputs show some tendency to approach the threshold of positive combinations, designed for wells 6 through 9.

Predicted secondary structures for each core type exhibit different tendencies. Thirty-three percent of the gates show no affinity for any unintended input, and their critical structures remain intact. Fifty percent of the gates show the previously noted tendency to yield a few pairs binding within the input regions, but not to the unintended inputs themselves. Eighteen percent of the gates, most of which were made of the E6 core, show varying affinities for inputs they were not designed for. Some exhibit partial input binding in the loop input region of the gates, and a few exhibit partial input binding to the stem portions of the gates. Generally, the secondary structures of the 8.17.1 cores have their stems (the critical region) remain intact in combination with incorrect inputs, whereas the E6 cores have far greater variability. The data supports the observation that the shorter critical region of E6 cores were less stable and prone to unintended bindings. Some of the high variability and unintended bindings occurred with gates composed of particularly rich GC content in the loop regions as well.

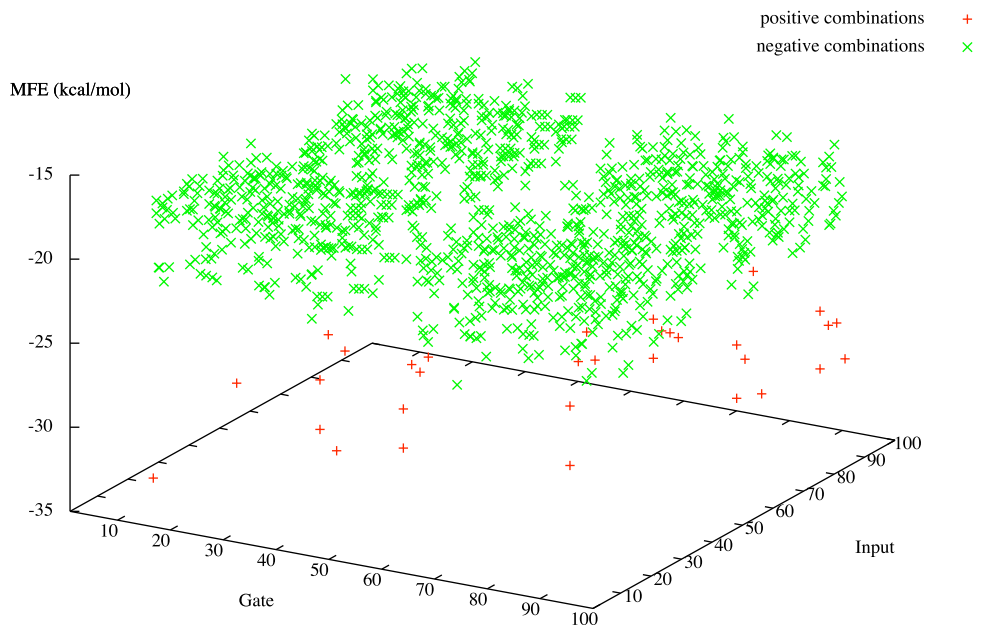

Figure 4: Distribution of minimum free energies for most probable deoxyribozyme gate and oligonucleotide input hybridization reactions.

3. Discussion

During the MAYA-II development time, the design process only allowed for limited assessment of the system as a whole. Gates were manually designed by modeling an individual deoxyribozyme structure with 15 nt sequences of interest inserted into one or more input binding regions depending on gate type. Free energy of one or more input bindings had never before been determined, nor was the system evaluated in terms of all potential reactions other than bioinformatics-based filtering to ensure sequence separation. During the experimentation process, time-consuming effort was expended on adjusting gate concentrations, and making small incremental changes to the basic gate designs in order to fine-tune the fluorescent signal output and avoid noisy combinations. These changes, as documented in [14], were based on laboratory experimentation alone and involved replacing inputs in a few cases, reversing loops, and adding or removing single nucleotides. Part of our reconstruction involves evaluating each of these incremental changes analytically, and then comparing directly with previous empirical observations, to see how well we can correlate model predictions to results, or, as the case may be, where the correlation fails. In both respects we obtain feedback not only on model predictiveness, but also on minute modifications of relatively simple reactions in analytical terms.

These critically important, but essentially very fine details, are now incorporated into a high-throughput approach to deoxyribozyme gate design. Our process is general such that millions of sequences can be tested, not only for optimal individual gate folding structures, but also for targeted binding to desired input, and lack of binding to other inputs that might be present in the mix. Moreover, our evaluation framework is modular, and designed to accommodate additional non-thermodynamic models which we feel will augment understanding of the underlying physical processes at work. We are actively working to tie new modeling results to empirical evidence. To the best of our knowledge, this approach has not been undertaken elsewhere and represents a verifiable, concrete path towards large-scale whole-system design. We view Pyxis as a dynamic tool, which can and should grow as different designs are proposed and evaluated. MAYA-II provides an excellent starting point and opportunity for deep probing, which can

now be fully captured and used in subsequent projects to shift the fine tuning more towards silicon testing.

We have answered here several basic questions for the simplest of MAYA-II gates. *Yes* gates fold properly while exhibiting some variable binding tendencies within loop regions. The MAYA input binding regions were pre-designed to provide open loop structures. However we expect collection of this loop secondary structure information to be useful in future gate designs, since it is expected that loop regions showing significant secondary structure would hamper input binding. Input secondary structure is a necessary parameter to be assessed in future input/gate designs, and future modeling will hopefully determine maximum binding parameters of successful gates.

Inputs preferentially bind to gates they were designed for. We saw that E6 gates had more variability in folding structures in the case of negative input binding, compared to the 8.17.1 core. This is most likely due to the smaller stem required by the E6 gate, which makes it functionally harder to design for. Model results show variability in binding with negative inputs which corresponds to experimental data generated during the original MAYA-II construction in two ways. Firstly, E6 gates required more manipulation of the stem region, as well as higher concentrations to obtain perfect digital behavior. Secondly, occasionally spurious signaling was noted early on after input introduction. The false signals tended to die down, and were later overtaken by the intended signal. Our current modeling suggest that use of the E6 cores may have been partly responsible for some of this behavior. The extra three nucleotides in the 8.17.1 core yielded reliable results for both positive and negative combinations. We expect that careful examination of the rest of the gates in various combination contexts, as well as gate designs that were rejected, will produce further inferences that can actively inform the design process. This work serves purpose in developing gates which can be pulled off the shelf and readily combined to suit logic goals; in essence it enables the development of a reusable oligonucleotide library.

3.1 Assessing modeling approaches

Historically, the basic design problem for molecular computing with nucleic acids has centered around sequence selection to achieve a particular target secondary structure. Design optimization methodologies dictate maximal probability of target structure stability based on energy considerations. Considering minimum free energy predicted structures or complexes is the necessary first step for thermodynamic evaluation of arbitrary collections of deoxyribozymes and oligonucleotides interactions. The next step we will take, however, is consideration of the remaining mass of possible, less probable complex formations, and determination of how much separation exists in terms of free energy compared to the predicted MFE complex. The MFE complex could in reality be dominated by the remaining possible complex formations, any one of which might not lead to the desired reaction, either in the positive or negative sense. We note that theoretical models such as DNA code words and strand algebra [2, 4, 9] which strive to formally codify notions of biological function and molecular computing strictly based on Watson-Crick base-pairing, do not capture the underlying jostling between competing structures which must be accounted for in the laboratory to actually make things work.

Outside of thermodynamics there are other modeling approaches and optimization possibilities, some or all of which can be codified into design rules. Strand secondary structure specific considerations [5] are GC content, Shannon entropy, and symmetry minimization. Strand tertiary structure considerations include DNA curvature, and base pairing and stacking where the nucleic acid geometries [8] are taken into account. Finally, the chemical kinetics of folding, cleavage, and hybridization [7] dictate how fast or slow reactions occur. We consider this aspect crucial since although energy considerations may yield the most stable structure or hybridization complex, in solution a faster reaction can deplete reactants and minimize formation of the predicted most stable one. Ultimately, our goal is to infer a weighting scheme for all factors determined to be the most relevant over the whole of a system, rather than relying on energy considerations alone and considering a system as a sum of individual non-interdependent physical

effects. Exploitation of this understanding suggests alternative control mechanisms as well where the energy landscape is intentionally kept highly variable, binding affinities are calibrated with deliberately inserted mismatches, or faster reactions are introduced to control the amount of products available for later reactions.

3.2 Acknowledgements

This material is based upon work supported by the National Science Foundation under NSF Grants 0829881 and 0829793. MLF is grateful for support from Elsa García and the UNM Intel Minority Engineering Scholarship.

References

1. Adleman, L. (1994) Molecular computation of solutions to combinatorial problems. *Science, 266*,1021-1024.
2. Andronescu, M., Dees, D., Slaybaugh, L., Zhao, Y., Condon, A., Cohen, B., Skiena, S. (2002) Algorithms for testing that sets of DNA words concatenate without secondary structure, *Lecture Notes in Computer Science, Revised Papers from the 8th International Workshop on DNA Based Computers: DNA Computing*, *Volume 2568*, 182-195.
3. Benenson, Y., Gil, B., Ben-Dor, U., Adar, R., Shapiro, E. (2004) An autonomous molecular computer for logical control of gene expression, *Nature, 429*, 423-429.
4. Cardelli, L. (2009) Strand algebras for DNA computing, unpublished.
5. Dirks, R., Lin, M., Winfree, E., and Pierce, N. (2004) Paradigms for Computational Nucleic Acid Design, *Nucleic Acids Research, Volume 32, Number 4*, 1392-1403.
6. Dirks, R., Bois, J., Schaeffer, J., Winfree, E., Pierce, N. (2007) Thermodynamic Analysis of Interacting Nucleic Acid Strands, *Society for Industrial and Applied Mathematics, Volume 49, Number 1*, 65-88.
7. Flamm, C., Hofacker, I. (2008) Beyond energy minimization: approaches to the kinetic folding of RNA, *Chemical Monthly, Volume 139, Number 4*, 447-457.
8. Harvey, S., Wang, C., Teletchea, S., Lavery, R. (2003) Motifs in Nucleic Acids: Molecular Mechanics Restraints for Base Pairing and Base Stacking, *Journal of Computational Chemistry, Volume 24, Number 1*,1-9.
9. Heitsch, C., Condon, A., Hoos, H. (2002) From RNA Secondary Structure to Coding Theory: A Combinatorial Approach, *Lecture Notes in Computer Science, Revised Papers from the 8th International Workshop on DNA Based Computers: DNA Computing, Volume 2568*, 215-228.
10. Hinsen, K. (2000) The Molecular Modeling Toolkit: A New Approach to Molecular Simulations, *Journal of Computational Chemistry,Volume 21, Number 2*, 79-85.
11. Huang, Y., Kowalski, D. (2003) WEB_THERMODYN: sequence analysis software for profiling DNA helical stability, *Nucleic Acids Research, Volume 31, Number 13*, 3819-3821.
12. Lafontaine, I., Lavery, R. (2000) Optimization of Nucleic Acid Sequences, *Biophysical Journal, Volume 79*, 680-685,
13. Langtangen, H. (2004) Python Scripting for Computational Science, Springer.
14. Macdonald, J., Yang, L., Sutovic, M., Lederman, H., Pendri, K. Lu, W., Andrews, B., Stefanovic, D., Stojanovic, M. (2006) Medium Scale Integration of Molecular Logic Gates in an Automaton, *Nano Letters, Volume 6, Number 11*, 2598-2603.
15. Markham, N. R. & Zuker, M. (2005) DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res., 33, W577-W581.
16. Privman, V., Strack, G., Solenov, D., Pita, M., Katz, E. (2008) Optimization of Enzymatic Biochemical Logic for Noise Reduction and Scalability: How Many Biocomputing Gates Can Be Interconnected in a Circuit?, *Journal of Physical Chemistry B,Volume 112, Number 37*, 11777-11784.
17. Stojanovic, M., Stefanovic, D. (2003) A deoxyribozyme-based molecular automaton, *Nature Biotechnology, Volume 21, Number 9*, 1069-1074.

18. Zuker, M. (2003) Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction, *Nucleic Acids Research, Volume 31, Number 13*, 3406-3415.