INCORPORATION OF PHYSICO-CHEMICAL PARAMETERS INTO DESIGN OF OLIGO PROBES FOR MICROARRAY EXPERIMENTS

Vladyslava G. Ratushna

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Department of Biological Sciences

Dr. Cynthia J. Gibas
Dr. Stephen M. Boyle
Dr. Jennifer W. Weller
Dr. Stephen B. Melville
Dr. Iuliana M. Lazar

May 6, 2005
Blacksburg, Virginia

Keywords: oligomer, probe, microarray, design, target, secondary, structure, *Brucella*, diagnostics

Incorporation of Physico-Chemical Parameters Into Design of Oligo Probes for Microarray Experiments

Vladyslava G. Ratushna

(ABSTRACT)

Microarrays containing long oligonucleotides provide sensitive and specific detection of gene expression and are becoming a popular experimental platform. In the process of designing an oligonucleotide microarray for *Brucella*, we optimized the overall design of the array and created probes to distinguish among the known *Brucella* species. A 3-way genome comparison identified a set of genes which occur uniquely in only one or two of the sequenced *Brucella* genomes. Reverse transcriptase PCR assays of over one hundred unique and pairwise-differential regions identified in *Brucella* revealed several groups of genes that are transcribed *in vivo* with potential significance for virulence. The structural and thermodynamic properties of a set of 70mer oligonucleotide probes for a combined *B. abortus, B. melitensis* and *B. suis* microarray were modeled to help perform quantitative interpretation of the microarray data. Prediction and thermodynamic analysis of secondary structure formation in a genome-wide set of transcripts from *Brucella suis 1330* demonstrated that properties of the target molecule have the potential to strongly influence the rate and extent of hybridization between transcript and an oligonucleotide probe in a microarray experiment. Despite relatively high hybridization temperatures used in the modeling process, parts of the target molecules are predicted to be inaccessible to intermolecular hybridization due to the formation of stable intramolecular secondary structure. Features in the *Brucella* genomes with potential diagnostic use were identified, and the extent to which target secondary structure, a molecular property which is not considered in the array design process, may influence the quality of results was characterized.

Dedicated To

Vladimir B. Gantovnik
and the ten ferrets,

who are my everyday joy and happiness.

ACKNOWLEDGEMENTS

I would like to thank all the members of my graduate committee ***Dr. Cynthia J. Gibas, Dr. Stephen M. Boyle, Dr. Jennifer W. Weller, Dr. Stephen B. Melville*** and ***Dr. Iuliana M. Lazar*** for scientific mentoring, and support during the work on this project. Thank you for sharing with me your extraordinary knowledge of micro- and molecular biology, biophysics and bioinformatics. Thank you for the access to computer and laboratory equipment and resources, as well as the financial support. It was a great pleasure to work under your guidance.

I would like to give special thanks to ***Dr. Betsy Bricker*** and ***Dr. Shirley Halling*** of the National Animal Disease Center (Ames, Iowa) for generously supplying the methanol killed *Brucella* biovars; ***Dr. Nammalwar Sriranganathan*** for the help in interpretation of the experimental results; ***David Sturgill*** for the identification of the differential regions in *Brucella*; and to ***Sheela Ramamoorthy*** and ***Sherry Poff*** for help with RT-PCR of the differential regions in *Brucella.*

I am especially grateful to my academic advisor ***Dr. Cynthia J. Gibas*** not only for the scientific guidance, but also for friendship, patience and care. You have created an atmosphere of warmth and kindness in our little bioinformatics lab, and I am truly happy to be a part of it.

I am also grateful to ***my friends, colleagues,*** and ***office mates*** at Virginia Tech for support and all of the great time that we had together.

Finally, I would like to thank ***my family*** for the many of years of love and support.

TABLE OF CONTENTS

LISTING OF FIGURES

LISTING OF TABLES

DESCRIPTION

The DNA microarray is a high throughput biotechnology method, which allows detection of thousands of genes in a single parallel experiment. The most common application is for gene expression measurements, although various types of comparative genome experiments are increasingly reported in the scientific literature. As the overall cost of microarrays decreases, use of the technique is becoming a standard rather than an extraordinary method. Some emerging applications of this technology are diagnostic assays and diagnostic comparative assays of reference pathogens to disease-causing isolates. Microarray experiments developed under two distinct design methods: in one the attached DNA molecule was a cloned cDNA fragment, which were deposited non-covalently on a slide surface. This format is still in use in many facilities. Affymetrix was the first company to manufacture the other major type of array, in this case called 'chips', on which a set of short oligonucleotide probes, covalently attached to the surface, represent parts of each target gene. These are the most commonly used platform in clinical studies. Affymetrix probes are only 25 nucleotides in length and there is some evidence [1, 2] that more sensitive microarray data can be obtained from array experiments which use custom-designed synthetic oligomer DNA probes, ranging from 50 to 70 nucleotides in length, with not too great a penalty in specificity. 70mer probes have been shown to have sensitivity similar to that of cDNA probes [3]. While sensitivity decreases with probe length [4-6], oligonucleotides shorter than 70 nucleotides are often used, and the greatest loss in sensitivity does not occur until oligos below 35 nucleotides in length are used [7]. These short synthetic oligonucleotides are designed using extensive computational analysis methods, including components that account for sequence identity, GC content, self-complementarity, Tm and other properties. The chief design issue addressed by most methods is to ensure that a probe is specifically and uniquely complementary to a particular segment of open reading frames (ORFs) in the genomes while having fairly uniform thermodynamic properties. The presence of known sequence variants must be accommodated to avoid confounding the signal. The discriminatory of any array design will always be subject to the experimental inputs, so a mixed genome sample or a related but non-reference sample to that used in the design should always induce a cautious approach to the data analysis. That is, the intent may be to use the array to detect specific cDNA molecules or mRNA transcripts in a mixture and differentiate between several related genomes, as well as study host-pathogen interactions, but all inputs must be examined before solid conclusions are drawn.

Probe molecules are either synthesized *in situ* or are printed to the microarray slide, and are either non-specifically cross-linked to the surface or are attached specifically using poly-Lysine linkers. The target molecules (in expression experiments these are most often fluorescently labeled cDNA molecules and sometimes cRNAs) hybridize in solution in a reversible bimolecular reaction to the probe oligomers, ultimately binding in stable double helices with their specific probes. The probe-target on-chip hybridization rate and stability are affected by several physico-chemical factors: hybridization temperature, probe length and specificity, G/C percentage of the probe sequence, probe location with respect to the 3'-end of the transcript, presence of stable secondary structure in the probe and/or target sequence and the probe-target hybridization kinetics. Most of these factors (excluding probably only the last two) are currently included as probe quality parameters in microarray probe design software packages. However, there are additional physico-chemical factors able to affect probe-target hybridization, which are so far very poorly incorporated in the experimental design considerations. In particular, the presence of the complex intramolecular bonding in the target molecules has been completely ignored. We have shown that these structures can be so stable that they neither unfold completely when the target mixture is heated to 65˚C in 1MNa$^+$, which is greater than the common microarray hybridization temperature, nor are they eliminated when the target molecules are sheared down to 50 nucleotides long fragments.

The long term goal of this research is to apply the oligo microarray technology to: a) the diagnosis of bacterial pathogens; b) compare gene expression in closely related bacterial species; and c) clarify the mechanisms of their host-pathogen interaction. This requires the design of species-specific probes capable of recognizing a transcript in a quantitative manner, with minimal interference from a background hybridization to other transcripts including those from the host, including mismatched binding to homologous target molecules. One aspect of designing a sensitive and quantitative diagnostic array is to eliminate those sites between probes and targets having a high potential for molecular interactions that may interfere with the intended hybridization and mask true differences in transcription levels. The investigation concerned the optimization of probe-target interactions on the oligo microarray with respect to the effect of stable secondary structures in the target and its amelioration the raise of hybridization temperature and decrease in the target length due to shearing. Our findings provide the background for establishing new criteria for the microarray probe selection.

OBJECTIVES

1. Use a widely accepted nucleic acid structure modeling software application [8, 9] to predict the most stable secondary structure in target molecules under experimental conditions. Model stable secondary structure formation in both cDNA and cRNA targets at temperatures related to the actual hybridization conditions (solvent characteristics and temperatures) used for performing the microarray experiments, and include a variable length of sheared fragments up to full length transcripts.

2. Analyze molecular properties of the oligonucleotide probes present in the current comparative genomic microarray for the three *Brucella* species: *B. abortus, B. melitensis* and *B. suis*. These properties include the probe sequence specificity, melting temperature of the probe secondary structure formation, presence of the polynucleotide stretches in the probe sequence, change in a Gibbs free energy, enthalpy and entropy on the probe secondary structure formation and the presence of mismatches if the probe is used for *B. abortus* and *B. melitensis* microarray experiments.

3. Compare the target and probe data from the existing *Brucella* array to determine which combinations have incompatible properties for complementary binding to occur.

4. Identify features (diagnostic signatures capable of producing a strong signal on probe-target hybridization) that will differentiate the three *Brucella* species to clarify and validate the RT-PCR results obtained earlier, and to compare the quality of the oligo probes designed using two different types of software either incorporating or skipping criteria based on models of the probe secondary structure formation.

1. INTRODUCTION

The first microarray experiment was reported in 1995 [10]. Early microarrays had full-length cDNA (copy DNA) molecules spotted onto an array surface. Today hundreds of microarray experiments are described in the scientific literature, many of which utilize synthetic oligonucleotide (oligo) probes 50 to 70 nucleotides in length [4, 5]. Oligo probe microarrays have a lot of advantages compared to cDNA microarrays: narrow Tm range and control of false hybridization, as well as means for detection of poor probes and problematic sequence regions. Oligonucleotide-based microarrays tend to yield more reproducible data and may now be overall less expensive than the cDNA microarrays. However, their use has been generally limited to those organisms having fully or extensively sequenced genomes, since synthesis of probes requires a complete knowledge of the sequence information whereas a cDNA may be attached whether or not one has full knowledge of its sequence. Until recently the cost of synthesis masks was such that the cost could only be justified for model organism arrays. With the advent of configurable masks this barrier has been eliminated and increasing numbers of oligo arrays are manufactured for less studied genomes (for example the grape and iceplant arrays manufactured by Nimblegen).

Existing methods for designing oligonucleotide microarrays primarily focus on sequence uniqueness of the target interaction site, secondarily on crude sequence composition characteristics and do not account for a number of biophysical characteristics of the probe and target molecules, which affect the hybridization behavior of these molecules. Our hypothesis is that ignoring these well-known effects in the design and analysis steps has led to the observation that array results are ‘only semi-quantitative’, useful for understanding trends and large changes but too inexact to be good diagnostic support. This narrowness means that many microarray experimental designs do not provide the sensitivity and specificity needed for diagnostic arrays. Fifteen to 30% of the probes initially selected by commercially available microarray design software packages fail to recognize their specific targets on the chip due to the limitations of probe design protocols (personal communication with Dr. Jennifer W. Weller, GMU).

1.1 BIOPHYSICAL CHARACTERISTICS THAT AFFECT PROBE-TARGET HYBRIDIZATION

The following biophysical characteristics are expected to affect the probe-target hybridization and hinder the ability to quantify microarray data. The characteristics are used by some of the available microarray probe design software as the criteria for the oligo probe quality. They are *often represented as user-adjustable characteristics, which can be either relaxed or narrowed depending on the needs of the particular microarray experiment.*

1.1.1 Sequence Specificity

A good oligo probe has a highly specific sequence, which can unambiguously hybridize to the expected transcript from a pool of target molecules under the experimental conditions. Probe specificity may be reduced by such factors as mismatches, polynucleotide stretches, presence of stable secondary structures on either probe or target molecules, etc. For example, long GC regions within the cDNA sequence may result in probe annealing to non-homologous sites on the target molecules. Current microarray probe design programs identify unique regions in the target using sequence comparison methods, and then search for the place suitable for probe binding based on other criteria. For example, the OligoArray [11] software reads the match length and percentage of identity values in the Blast output and compares them to the threshold parameters to match the sequence specificity level, required by a user.

1.1.2 Probe Length

Use of oligonucleotides that are much shorter than the corresponding cDNA is one of the major advantages of the oligo probe microarrays. This allows the user to avoid regions of extensive homology and stable secondary structure and still pick highly specific probes. A smaller probe length also reduces the reagent-per-microarray cost of the experiment for some manufacturing schemes. Several non-overlapping probes spaced across the target sequence may be used as a type of within-array transcript replicate in one sample. To use the information in such a way it is a requirement that the probes be optimized to have uniform physico-chemical properties. For most gene expression experiments a number of experiments have demonstrated that the optimum oligo probe length is between 50 and 70 nucleotides, although some arrays are designed with probes 27-35 nucleotides long [3, 4, 12]. The extended length delivers higher sensitivity compared to traditional shorter length probes. It also allows for the design of microarrays with fewer, more selective probes. Therefore, the oligo length is often given as an adjustable parameter in oligo design software.

1.1.3 Probe Secondary Structure

It is intuitively obvious that the longer the region of homology the more likely that it is to be impinged upon by a region of stable secondary structure, in either the probe or target molecule.

Usage of short synthetic probes allows avoiding the regions in the target sequence having high potential for secondary structures. The presence or absence of competing stable structure in a particular place on either the probe or the target DNA molecule is a crucial factor in success or failure of the entire microarray experiment. Therefore, it is important to predict the formation of such structures across the entire flexible and sticky nucleotide strand in order to obtain reliable data. At least one probe design software, called OligoArray [11], explicitly computes the melting temperature of the most stable structure forms in the probe molecule to filter out the secondary structure in the probes.

1.1.4 Mismatches

The degree of mismatch that must exist to be able to neglect cross-hybridization is a complex function of length, distribution and structure for all partially complementary sequences. It represents a range of possible binding events, and is a time-dependent phenomenon. Since the off-reaction is the slow step, kinetics becomes very important in the discrimination of mismatches.

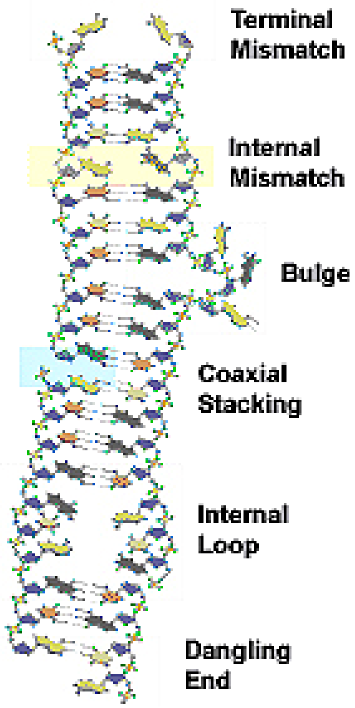


Figure 1.1.4
DNA structural motifs
Image compliments of DNA Software, Inc.

There are several common structural motifs that may cause false positive results: terminal mismatch, internal mismatch, bulge, coaxial stacking, internal loop and dangling ends. Most of the available software uses an incomplete nearest neighbor parameters set, which results in a calculation of the melting temperature that can be incorrect by more that 6ºC. Oligonucleotide probe hybridization to a long target DNA involves not only base pairing, but also two dangling-end contributions. Depending on the sequence, two dangling ends can affect probe–target hybridization by as much as +0.96 to –1.92 kcal/mol. By comparison, the stabilization conferred by addition of one Watson–Crick base pair to a probe–target duplex ranges from –0.58 (TA/AT) to –2.24 kcal/mol (GC/CG) [13]. The only software that accounts for the dangling end effect is under development by J. SantaLucia [14].

Design and analysis tools booleanize both parts of the microarray process with an inevitable loss of information.

1.1.5 G/C Content

G/C percentage is an important oligo probe characteristic. It gives a crude approximation of the oligonucleotide-target complete hybrid stability, expressed as the melting temperature, and is based upon rather simple algorithms that are trivial to calculate, therefore, it is almost always included in microarray probe design software. The G/C content above 50% is not desirable, because combined with the other factors it may lead to non-specific probe-target hybridization. The main problem with the short probes with either GC or AT rich sequences is that they limit the design space for the other oligos. For long hybrids it may be difficult to attain the level of stringency desired, i.e. to limit the stability of mismatches, when GC content is high. For the species with high G/C percentage it may become a limiting criterion for the probe selection, and will restrict the oligo length down to 50-60 nt.

The following are several physico-chemical conditions which are not directly included in the process of microarray oligo design as it is currently implemented, but are important criteria for the correct evaluation of the probe behavior on a chip.

1.1.6 Hybridization Kinetics

The probe-target interactions on the microchip are one part of a complex chemical equilibrium process. In the absence of competing reactions (such as internal bonding) a probe and target are part of a simple second-order chemical equilibrium whose end-point is determined by concentration and whose rate at achieving that end-point depends on the joint concentrations and the temperature/solvent conditions. The probe is intended to be present in sufficiently high concentration that the overall reaction is pseudo-first order. Thus, for some concentrations of reactants, the amounts of the probe/target hybrid that is formed can be predicted, or, more important for microarray experiments, from the amount of hybrid detected the starting concentration of the reactants can be predicted, in particular of the target since the probe is present in vast excess. While short nucleic acid hybrids do not have a reaction intermediate (transition complex) there are a number of factors that decrease the accuracy of the estimate of the concentration: imperfect probe density on a microchip, trace reagents affecting hybridization in unexpected ways, interactions of target with the array surface or coating, imperfect temperature control, and so on. According to one mathematical model, developed for heterogeneous DNA-DNA hybridization, there are two different mechanisms by which targets can hybridize with the complementary probes: direct hybridization from the solution, and indirect hybridization of molecules that were first adsorbed nonspecifically to the array surface itself, and that subsequently diffused across the surface until coming into proximity with a probe. It was shown that nonspecific adsorption of single-stranded DNA on the surface followed by two-dimensional diffusion significantly enhances the overall hybridization rate [15].

This type of heterogeneous hybridization depends strongly on the rate constants for DNA adsorption/desorption in the non-probe-covered regions of the surface, the two-dimensional (2D) diffusion coefficient and the size of probes and targets. The diffusion of single stranded DNA is constantly interrupted by repeated association and dissociation with immobile oligonucleotide molecules. Experimental studies show that the hybridization efficiencies of 5'-end support-bound oligonucleotides are 75-80% for single-stranded oligonucleotide targets and 40-50% for the long double-stranded targets, respectively [16].

1.1.7 Ionic Strength

The presence of monovalent and divalent cations and other chemicals that affect the solvent dielectric constant are environmental properties that alter the dynamics of probe-target hybridization. For example, the addition of formamide reduces the dielectric constant and therefore the hydrogen bond energy of the base interactions, thereby decreasing the apparent (externally measured) most effective hybridization temperature. The ionic strength is am important factor in the stringency (specificity) of a hybridization solution, since phosphate neutralization is required for strands to approach to H-bond forming distances. There are a number of empirical correction equations, which account for DNA thermodynamics under particular reaction conditions of varying concentrations of sodium, magnesium, urea, DMSO and formamide, etc.

The known properties of the nucleic acids suggest that the presence of the stable secondary structures on the long target molecules will affect the probe hybridization on the microarray, however the consequences of the presence of these structures on quantitative approximations have not been explored.

1.1.8 Target Secondary Structure

It is possible that long target molecules of either cDNA or cRNA have a reduced binding affinity for microarray probes due to the formation of internal stable secondary structures that obscure the intended probe binding sites. Different intramolecular foldings such as hairpins and stacked regions may pre-empt base availability of target nucleotides, thus blocking regions of the long RNA (or cDNA) molecules from hybridizing to their intended probes. Our preliminary computational analysis indicates that raising the probe-target hybridization temperature and decreasing the overall length (reducing the available intramolecular structure domain) with shearing techniques decreases the frequency and stability of secondary structures, but does not eliminate the presence of short stable helices on the target molecule. These temperature resistant structures reduce the effective concentration of target available for binding the probe and thus lead to an inaccurate estimate of the concentration, or even the conclusion that the target is not present at all. An algorithm to identify and mask target regions having a strong propensity to form such resistance structures should be incorporated into microarray probe design software applications as an additional filter.

1.2 THE MODEL ORGANISM

An essential part of our project is an experimental investigation involving three bacterial species from the genus *Brucella: B. abortus, B. melitensis* and *B. suis*. *Brucella* is a facultative intracellular pathogen with an approximately 3 Mb genome, the total being divided between two chromosomes of sizes of 1.85 Mb and 1.35 Mb. Human brucellosis is quite common but often not diagnosed [17]. There are six recognized *Brucella* species that differ in their host preference. *B. abortus* preferentially infects cattle, *B. melitensis* infects sheep and goats, and *B. suis* infects pigs. All three of these species and *B. canis* can infect humans, although *B. melitensis* is associated with the most serious human infections. The *Brucellae* are grouped with the α-proteobacteria and are related to other cell-associated parasites of plants and animals [18]. The true pattern of *Brucella* intracellular survival and proliferation, and the reasons for the different virulence patterns among the species, are not conclusively known. Human vaccines against *Brucellae* are not currently available [19].

The genomes of *B. melitensis*, *B. suis* and now *B. abortus* are completely sequenced and publicly [20-22] Development of a comparative expression microarray for these three *Brucella* species will allow an investigation of the different aspects of functional genomics of *Brucellae,* such as: Which genes are involved when *B. melitensis* invades and multiplies in each of its diverse hosts? How useful are microarrays as a platform to elucidate the differences in host-preference among the three *Brucella* species? An important long-range goal of our collaborators is the development of a diagnostic miniarray containing probes both unique to particular *Brucella* species and common within the *Brucellae*, a tool that will have direct application to identification of the source of a *Brucella* spp. associated with an infection.

2. RESULTS

2.1 TARGET SECONDARY STRUCTURE MODELING

Secondary structure in the target is a property not usually considered in software applications for design of 'optimal' oligonucleotide probes. It is frequently assumed that eliminating self-complementarity, or screening for secondary structure in the probe, is sufficient to avoid interference with hybridization by stable secondary structures in the probe binding site of the target. Prediction and thermodynamic analysis of secondary structure formation in a genome-wide set of transcripts from Brucella suis 1330 *demonstrates that the properties of the target molecule have the potential to strongly influence the rate and extent of hybridization between transcript and tethered oligonucleotide probe in a microarray experiment* [23].

2.1.1 Background

Sequence-specific hybridization of a long single-stranded labeled DNA or RNA target molecule to a shorter oligonucleotide probe is the basis of gene expression microarray experiments. In this type of microarray experiment, gene specific *probe* molecules are either synthesized *in situ* or are printed to the microarray slide, and are either non-specifically cross-linked to the surface or are attached specifically using a method such as poly-Lysine linkers. *Target* molecules (most often fluorescently labeled cDNA molecules, although cRNAs are used in some protocols) hybridize to probes on the chip until most of them find and form stable double helices with their specific probes. Transcript abundance is estimated by the relative intensity of signal from each spot on the array. This interpretation of array data assumes that each hybridization reaction reaches equilibrium within the duration of the experiment.

There are three major types of DNA microarray platforms, which differ in the probe manufacture and attachment methods and the array size and chemistry. A cDNA microarray utilizes probes of up to several hundred base long PCR-generated amplicons that may be deposited or crosslinked to a coated glass slide surface [10]. Affymetrix microarrays [24] and similar *in situ* synthesized, covalently attached short probe designs use a set of 25-mer oligonucleotides complementary to sites, usually non-overlapping, distributed along the 500 nucleotides at the 3' end of the intended target molecule. Synthetic long-oligomer probe microarrays are used in a variety of commercial and custom settings. This class of microarrays utilizes chemically synthesized 35-70 nucleotide long probe [4-6] per each target sequence, deposited on a glass substrate that may or may not be covalently attached and that usually includes a moderately hydrophobic linker attaching it to the slide coating. Some recent studies indicate that synthetic oligomers of up to 150 nucleotides may be desirable for assessing transcript abundance [12]. The optimal probe length is directly related to the purpose of an experiment and the nature of the target mixture being assayed. In general, the use of synthetic oligomers has been shown to result in improved data quality [1, 2] relative to cDNA arrays, and 70mers have been shown to detect target with a sensitivity similar to that of full length cDNA probes [3]. Short probes have been preferred compared to the long ones, because they facilitate finding unique sequence matches while forming fewer, and less stable, hairpin structures and displaying more uniform hybridization behavior overall. However, the need for sensitivity and detection of transcripts in low copy number suggests the use of long-oligonucleotide arrays. In this study, we have modeled the accessibility of transcripts to hybridization with 70mer oligonucleotides.

In recent years a number of oligonucleotide design software packages have been published [11, 25-28]. Several factors are considered by almost all microarray design software packages: especially the sequence specificity of the probe-target interface and the overall balance of GC content across the array. A relatively uniform melting profile is generally achieved simply by selection of probes with similar GC content and length. However, some design methods explicitly compute the duplex melting temperature for each candidate probe-target pair and filter unique probes to find those which match a specified range of melting temperatures. Another biophysical criterion sometimes applied to probe selection process is elimination of probes having the ability to form stable intramolecular structures under the conditions of the experiment. This is usually done by eliminating regions of self-complementarity.

Very few of the available array design packages consider the possible structures of the transcript-derived molecules in the sample solution and their effect on the microarray quality. It has been shown that a 6 base long hairpin in a target will require a 600-fold excess of the complementary strand to partially displace the hairpin [29]. Since the target molecules are generally longer than probes and may have different backbone chemistry in the case of RNA, it is not sufficient to conclude that their behavior will mirror the behavior of the complementary probe. Prediction of secondary structure in a sample transcript using a standard secondary structure prediction algorithm for nucleic acids (Mfold) demonstrates that, although longer-range interactions are reduced at high temperatures, stable local structures persist in the transcript even at high salt concentration (Figure 2.1.1).

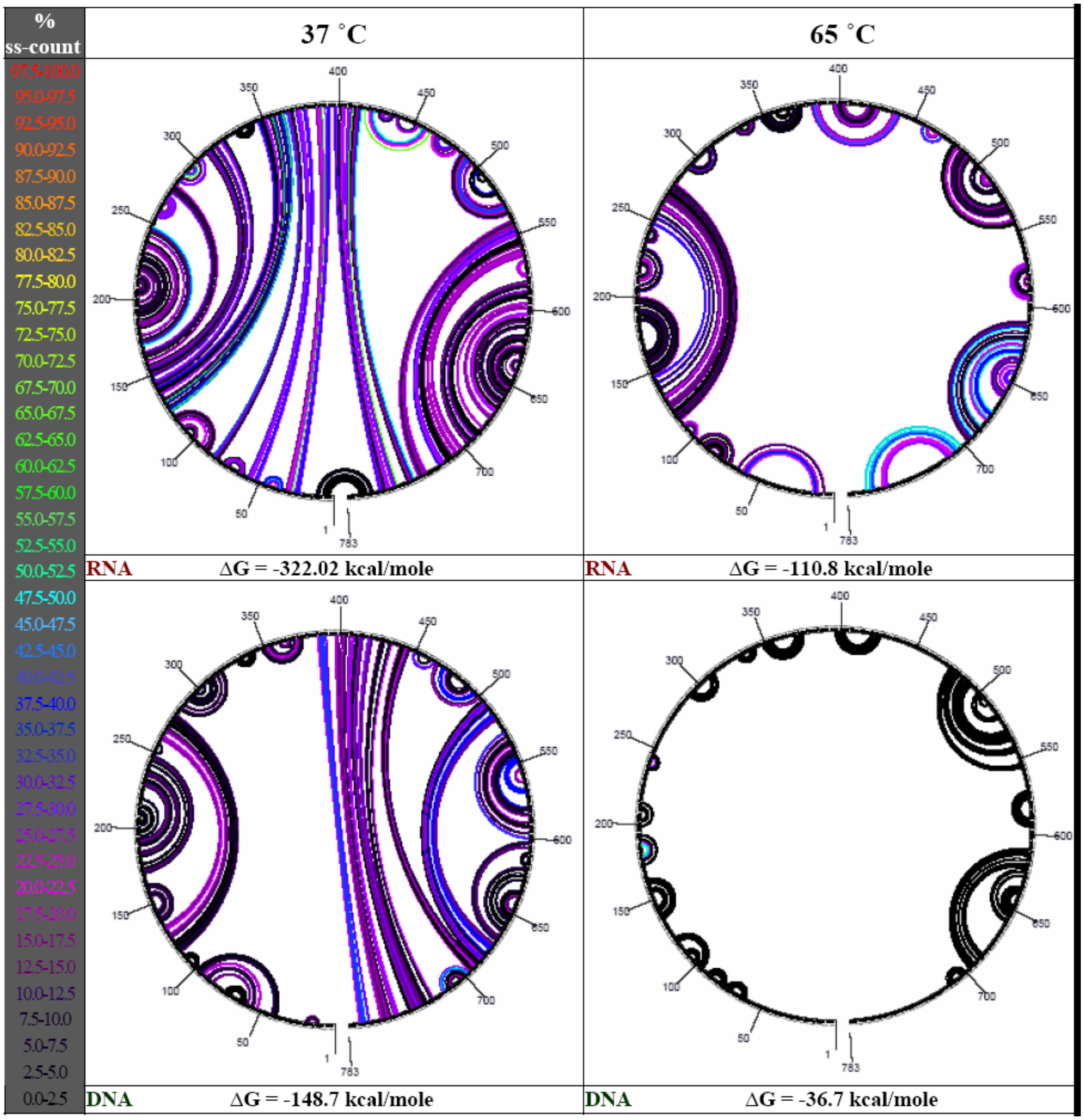


Figure 2.1.1 Secondary structure in a sample transcript. Circular diagrams of structure in a sample transcript (*moeB* homolog designated BR0004) from *Brucella suis*. Circular diagrams show hydrogen bonds between individual nucleotides, color-coded according to single-strandedness – the fraction of structures in which that bond is not present. Black bonds indicate 0% single-strandedness; red bonds indicate 100% single-strandedness.

Since they are unimolecular reactions intramolecular bonding within the target will occur considerably faster than diffusion-mediated bimolecular hybridization reactions. Therefore, the intramolecular structures are expected to be present and able to block the specific probe annealing sites on the target sequence in some cases, decreasing the available concentration for hybrid formation and thus resulting in misinterpretation of the signal obtained from the assay.

In order to estimate the abundance of stable secondary structure in long target molecules, and the impact of such structures on the analysis of microarray data, we have modeled secondary structure formation in mRNA transcripts of the intracellular pathogen *Brucella suis*. We have assessed the stability of structures formed in the transcript and the accessibility of the binding sites of optimal probes generated using commonly applied design criteria. We have also modeled the effects of random shearing of the full length target molecules on the abundance of secondary structure in selected targets.

2.1.2 Methods

Prediction and thermodynamic analysis of secondary structure was performed for all protein-coding gene transcripts from 3264 CDSs in the *Brucella suis 1330* genome. *Brucella suis* has a relatively high (57%) genomic GC content. It was chosen for this experiment due to availability of a custom synthetic oligomer microarray for this organism, developed by TIGR using standard oligo array design software, along with a set of unique probe sequences.

In order to determine whether GC-rich *Brucella* sequences form unusually stable structures we randomly picked and analyzed 50 gene coding sequences from a compositionally balanced genome of *Escherichia coli*, and 50 from the AT-rich genome of the nonpathogenic gram-positive bacterium *Lactococcus lactis* (35% genomic GC content). The *Brucella suis* genes ranged in length from 90 to 4,803 bp, with an average transcript length of 851 bp. The *E. coli* genes ranged in length from 140 to 2,660 bp, with an average transcript length of 792 bp. The range of GC content in the genes chosen was 37% to 57% with an average value of 50%, which is reasonably representative of the *E. coli* genome. The *L. lactis* genes chosen ranged in length from 140 to 2,730 bp, with an average transcript length of 765 bp, and ranged in GC content range from 30% to 42% with an average value of 35%.

2.1.3 Microarray design

70-mer probes for each *Brucella suis* target were custom designed (Dr. Stephen M. Boyle, Virginia Tech, personal communication) using ArrayOligoSelector (pick70) [25]. ArrayOligoSelector uses sequence uniqueness, self-complementarity, and sequence complexity as criteria but does not explicitly evaluate ΔG of secondary structure formation for the probe. 72% of the probes designed using this method were found to contain secondary structures with melting temperatures greater than 65°C, and 10% contained secondary structures with melting temperatures greater than 80°C. We evaluated the structural accessibility of the target probe-binding sites defined by the probes designed by pick70.

2.1.4 Secondary structure prediction

Probe and transcript secondary structure were predicted using the Mfold 3.1 software package [9, 30]. Mfold predicts the optimal folding of a nucleic acid sequence by energy minimization and can predict suboptimal foldings within a specified energy increment of the. Secondary structure in the single-stranded DNA and RNA target was modeled at a range of hybridization temperatures commonly used in microarray protocols: 37°C, 42°C, 52°C and 65°C. The modeling conditions were chosen to simulate a microarray experiment (1.0 M sodium concentration and no magnesium ion). The free energy increment for computing suboptimal foldings, ΔΔG, was set to 5% of the computed minimum free energy. The default values of the window parameters, which control the number of structures automatically computed by Mfold 3.1, were chosen based on the sequence length. Free energy changes on formation of secondary structure were extracted from the Mfold output.

2.1.5 Accessibility calculation

There are few experimental studies on accessibility of folded single-stranded DNA or RNA, and since the structure of such molecules in solution is very dynamic, each molecule is likely to exist in an ensemble of structures. Use of the Sfold server [31, 32], with batch jobs limited to 3500 bases, is not currently practical for a genome-scale survey of accessibility. Another approach to accessibility prediction is McCaskill's partition function approach [33] which can be used to compute base pair probabilities and summary pairing probability for any base. This approach is used in RNAfold [34], a component of the Vienna RNA package.

In this study, we use a simple fraction of predicted optimal and suboptimal structures in which a residue is found to be part of a single stranded structure, as computed by Mfold as an approximation of probability of single strandedness. Accessibility scores derived from Mfold predictions have been used in limited studies of RNA structure focused on ribozymes [35], antisense and siRNA targeting [35, 36] and have been shown to be experimentally predictive . We have chosen to use the Mfold, because its accessibility scores have been used with reasonable success to predict accessibility in the siRNA [37].

2.1.6 Shearing simulation

Random shearing of the target mixture is often used as a solution for the problem of target secondary structure. Shearing breaks the DNA or RNA molecule in random locations and gives rise to a mixture of fragments. We picked fragments of 200, 100, or 50 bases in length to simulate the effects of shearing on structure formation and stability in a transcript and used a 10 base sliding window approach. Secondary structure prediction for all fragments derived from every transcript in the *B. suis* genome is computationally intensive and produces an extremely large amount of output. Since our initial goal was to determine how much the method would affect the number and type of secondary structures probes would be expected to bind the shearing simulation was performed for fragments derived from the 300 bp Ure-1A gene of *B. suis*. Secondary structure and thermodynamics were computed for each of these fragments individually.

2.1.7 Results

Despite the relatively high hybridization temperatures and 1M monovalent salt hybridization conditions used for the modeling process, parts of the target molecules are likely to be inaccessible to hybridization due to the formation of stable secondary structure. Our *in silico* modeling shows that at 65°C, 28 ± 7% of the average cDNA target sequence is likely to be inaccessible to hybridization. The analysis of the specific binding sites of a set of 70mer probes designed for *Brucella* uses a freely available oligo design software package. 21 ± 13% of the nucleotides in each probe binding site are found within a double-stranded structure in over half of the folds predicted for the cDNA target at 65°C. The structures formed on the target are more stable and extensive when an RNA molecule is modeled rather than cDNA. When random shearing of the target is modeled for fragments of 200, 100 and 50 nt, an overall destabilization of secondary structure is predicted, but secondary structure does not disappear completely with shearing.

2.1.7.1 Extent and stability of target secondary structure

The results obtained for the complete genome-wide set of intact single-stranded DNA or RNA targets demonstrate that stable secondary structures are widespread in target mixtures from *Brucella suis* (Figure 2.1.7.1.a) and in randomly chosen transcripts from the genomes of *E. coli* and *L. lactis*. Figure 3 shows the ΔG of formation for the most stable predicted secondary structure of the full-length transcript, as a function of hybridization temperature.

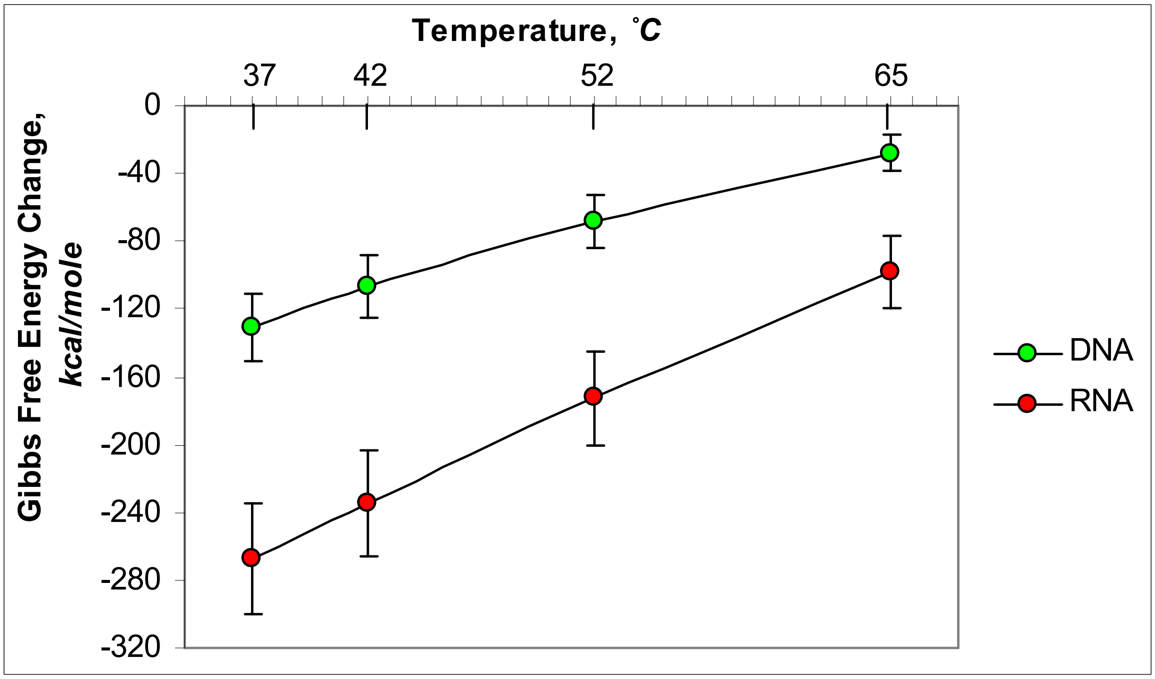


Figure 2.1.7.1.a Stability of transcript secondary structure in *Brucella suis.* Average free energy change on global secondary structure formation for *Brucella suis* targets, modeled as DNA or RNA. ΔG values are normalized to global mean target length.

The energies were normalized by computing a per-residue folding ΔG for each transcript and then multiplying that value by the global mean target length, for all transcripts considered from all organisms, of 851 bp. Average ΔG of secondary structure formation decreases with increasing temperature, but even at 65°C, the average ΔG of secondary structure formation for a full-length transcript is -98.2 kcal/mol for RNA and -27.9 kcal/mol for cDNA). This means that the transcript is quite stable in that structure and a considerable energy input will be required to displace or melt the remaining structure. The overall stability of the secondary structure formation from the high-GC genome of *B. suis* to the low-GC genome of *L. lactis* decreases. The average normalized ΔG of secondary structure formation for transcripts selected from the GC-balanced genome (*E. coli*) is near 70% of the average for *Brucella*, while the average ΔG for transcripts from the GC-poor genome (*L. lactis*) are even lower (30% at 52°C). However, even in the most GC-poor genome, stable target secondary structure in the single-stranded target is widespread.

Our results demonstrate that a significant fraction of nucleotide sites either single stranded DNA or RNA has stable secondary structure under the hybridization conditions used in oligonucleotide microarray experiments, and is relatively inaccessible for intermolecular interactions. Figure 2.1.7.1.b shows the percentage of nucleotides in a double-helical state in at least 50% of the secondary structure conformations predicted by Mfold, at various reaction temperatures. The fraction of all predicted optimal and suboptimal structures in which a nucleotide is found in a single-stranded conformation was used as a measure of accessibility.

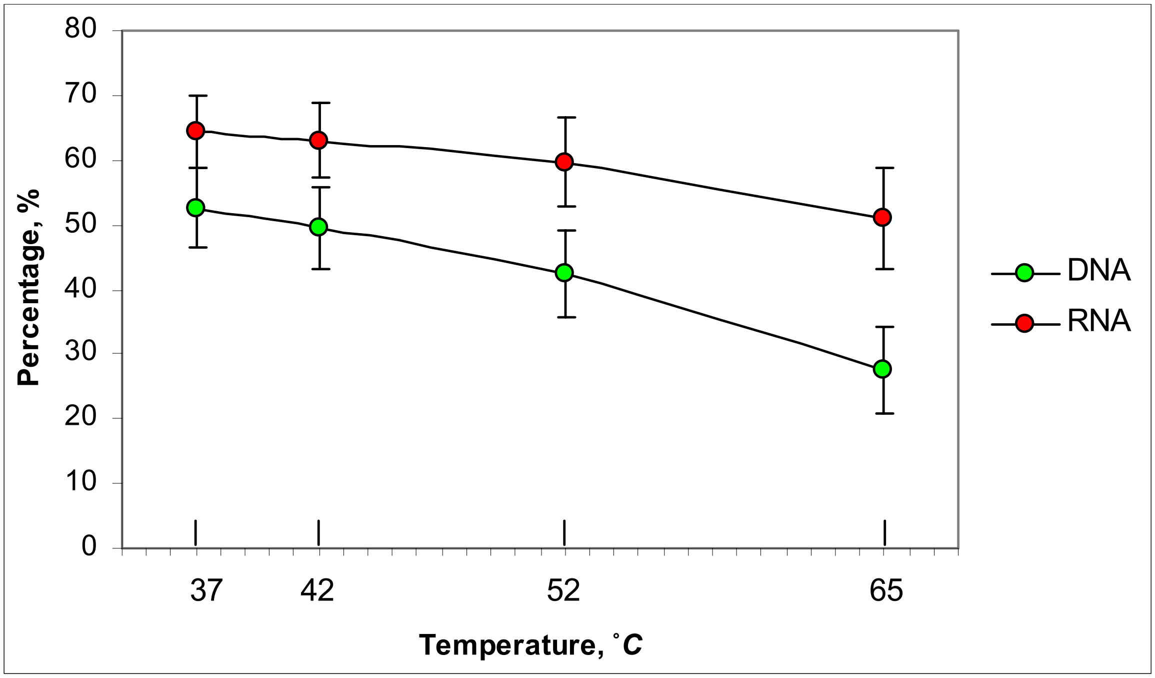


Figure 2.1.7.1.b Fractional accessibility of nucleotides in the target. Fraction of the complete transcript classified as inaccessible due to the presence of stable structure in >50% of predicted conformations. Data shown are for 37, 42, 52 and 65°C simulations in *Brucella suis*.

2.1.7.2 Extent and stability of sheared target secondary structure

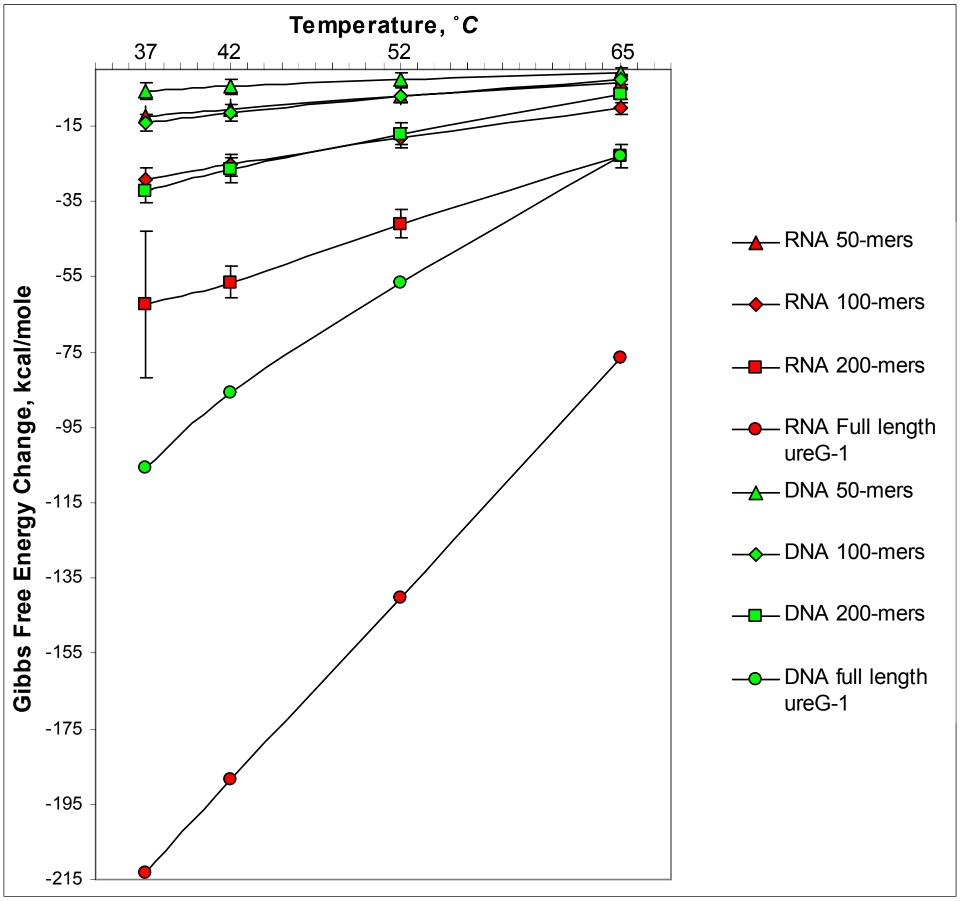


Figure 2.1.7.2 Stability of secondary structure in sheared fragments. Free energy change on secondary structure formation for the ureG-1 RNA transcript from *Brucella suis*. The transcript is modeled as sheared into fragments of length 200 nt, 100 nt or 50 nt; fragments are chosen starting at every 10th residue.

Figure 2.1.7.2 has a plot, which shows the average ΔG of structure formation for shearing simulation of the target molecule into overlapping 200, 100, and 50mer fragments. Shearing the target into smaller fragments destabilizes secondary structure, especially at very short fragment lengths. However, shearing does not eliminate completely the secondary structure, even in the shortest fragments examined. When a DNA target is modeled at 52°C, for example, the double stranded fraction decreases by only about 30% – from 41% to 29% – when the target is simulated as sheared into 50mer fragments. However, in hybridization experiments involving low copy number targets and longer oligos, creating extremely short target fragments may substantially reduce the signal by creating the fragments that can only partially match the probe.

2.1.7.3 Interference of secondary structure with the hybridization

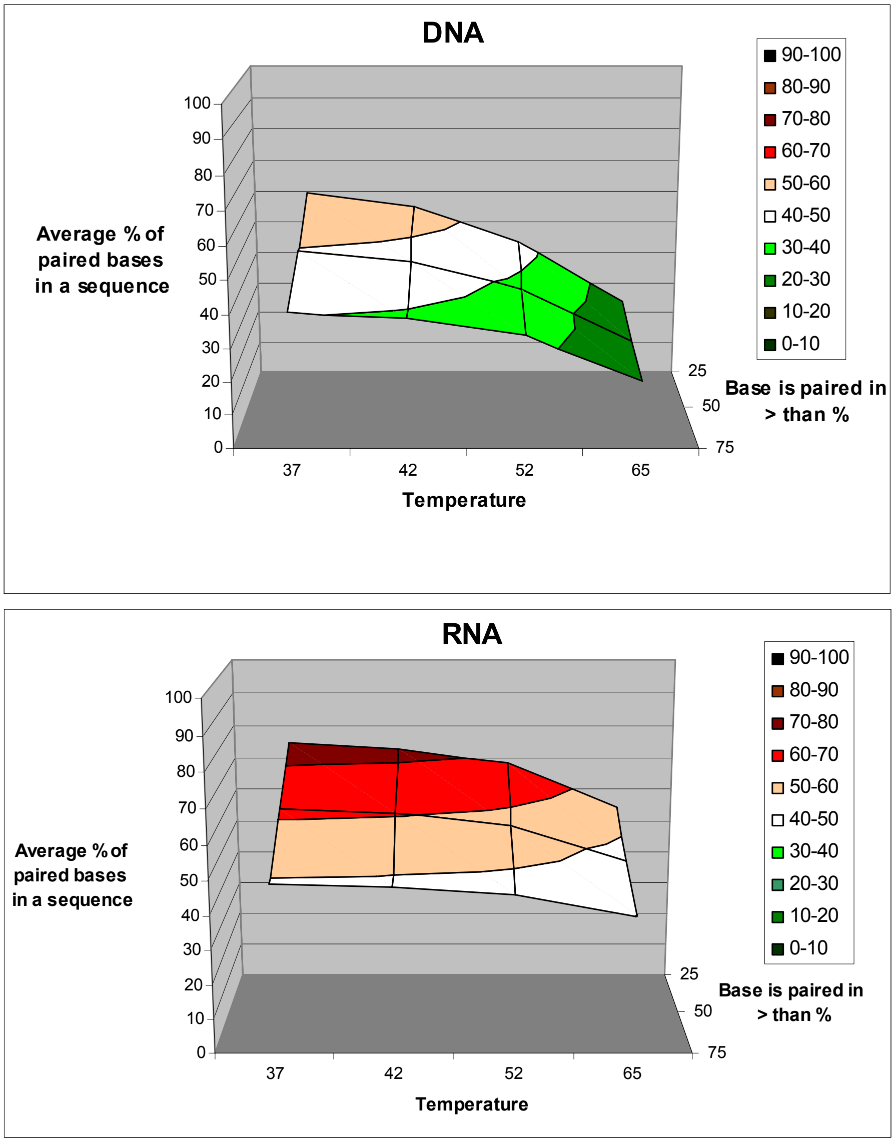


Figure 2.1.7.3.a Accessibility of the probe binding site. Fraction of the average probe binding site in the *Brucella* genomic array that is found to be inaccessible at 37°, 42°, 52° and 65°C, for DNA or RNA target. Inaccessible sites are defined here using three different cutoffs for the fraction of structures in which the site is base-paired: 25%, 50%, and 75%.

Figure 2.1.7.3.a shows the average percentage of nucleotides within a probe binding region in the target that are inaccessible, when different fractional accessibility cutoffs are used to classify the sites. For example, at 65°C and double-strandedness over 75% of optimal and suboptimal structures an average of 21 ± 13% of nucleotides in the probe binding region are found forming stable secondary structures at 65°C.

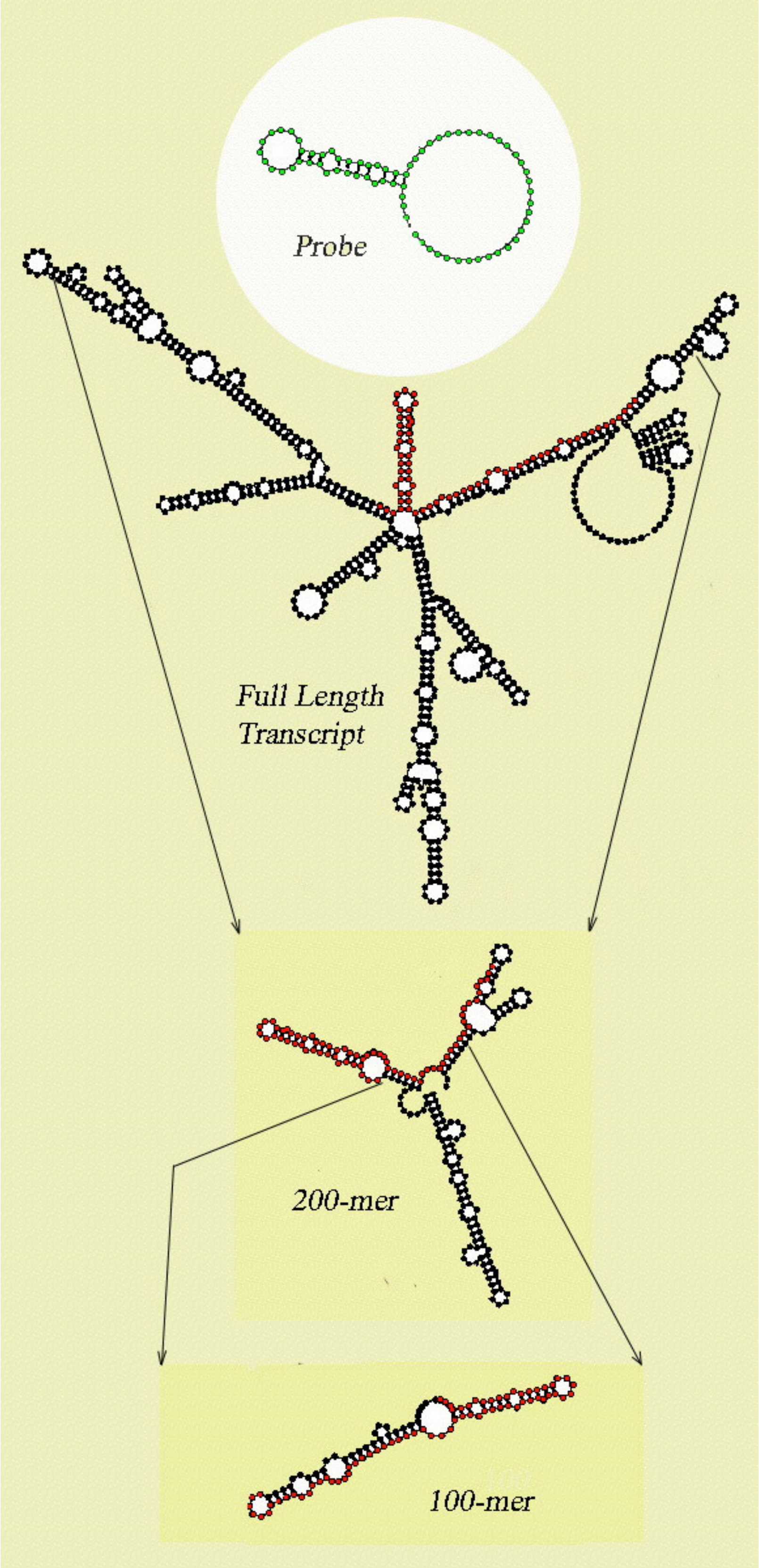


Figure 2.1.7.3.b Structure in a binding site – full length target and sheared fragments.

The position of a 70mer oligonucleotide probe (green) binding site (red dots) within a full-length optimal transcript structure, as well as examples of stable structure in 200mer and 100mer fragments which overlap the probe binding site. Corresponding ΔG values for these fragments modeled at 42° and 52°C are shown in Table 2.1.7.3.

Figure 2.1.7.3.b shows a representative transcript and the challenge it presents to hybridization when modeled as full-length cDNA and fragments of various lengths.

Table 2.1.7.3 Stability of a sample transcript – full length target and sheared fragments. Folding ΔG of target transcript and fragment molecules shown in Figure 8, at hybridization temperatures commonly used for long oligomer arrays.

Molecule	**ΔG, kcal/mole**			
	42°C		**52°C**	
	DNA	**RNA**	**DNA**	**RNA**
70-mer Probe	- 6.8	N/A	- 4.2	N/A
Full Length Target	- 85.9	- 188.4	- 56.6	- 140.2
200-mer sheared Target	- 25.5	- 58.6	-15.9	- 41.6
100-mer sheared Target	-14.2	- 25.7	-9.6	-18.0
50-mer sheared Target (not shown)	- 6.1	-10.5	- 4.2	-7.3

2.1.8 Discussion

Lack of bioinformatics tools that incorporate experimentally validated biophysical properties of nucleic acids as a criterion for synthetic oligomer probe design is a major challenge for microarray designers. We predict that the propensity of long single-stranded DNA or RNA molecules to form stable secondary structure will reduce the binding efficiency of microarray probes to their targets. 3-D structures such as hairpins and stacked regions have the potential to block regions of the target molecules from hybridizing to their intended probes. Prediction and thermodynamic analysis of secondary structure at a range of temperatures in full length target sequences, as well as in subsequences formed by *in silico* shearing, revealed the likely presence of stable secondary structures in both full-length target and sheared target mixtures. An increase in the hybridization temperature and more extensive shearing do not convert these structures completely to random coils. Such secondary structures are certain to compete with the sites on the target intended for annealing to the probe, which will result in misinterpretation of microarray data unless the effect is recognized and compensatory normalization is applied to the data.

2.1.8.1 Applying target secondary structure as a criterion in array design

The results from this modeling experiment demonstrate that elimination of probe secondary structure by avoiding self-complementarity is not by any means a transitive process for the target, by which we mean that it does not eliminate target secondary structure. Use of target secondary structure in addition, as an explicit microarray probe design criterion, will allow the researcher to mask and avoid the regions of the target sequence involved in very stable and ubiquitous secondary structure formation.

In this study the assigned accessibility scores to bases in the target sequence was based on the fraction of predicted structures, in which the residue is found in a single-stranded conformation. This approach equally weights each possible structure in the ensemble of optimal and suboptimal structures and considers bonds which form only in rare conformations equal to bonds which are present in the lowest-energy structure. The program Sfold [31, 32, 38] assigns a score for the residue accessibility based on an ensemble-weighted average of secondary structure. The program RNAfold [34], part of the Vienna RNA package, implements McCaskill's partition function approach [33] to calculate the pairing probabilities for each pair of bases in the sequence. Although it has been shown that all three software applications yield nearly identical results when predicting the binding states for molecules of known secondary structure [38], the use of the latter two software applications for solving the problem of the target secondary structure may produce somewhat different results than the Mfold application.

Mfold-based accessibility predictions for an individual transcript were compared to those generated by Sfold and RNAfold; and the difference in the average predicted accessibility over an entire transcript is small. The accessibility for the transcript of human ICAM-1, which has been mapped experimentally, was computed to determine its accessibility [39]. The average fractional accessibility derived from Mfold results is about 3–4% greater than that predicted by RNAfold or Sfold. Therefore use of this fractional accessibility measure will not impose an unnecessarily stringent constraint on the design process relative to other predictive approaches. The accessibility profiles calculated for ICAM-1 using each method are shown in Figure 2.1.8.1. In each section of the figure, antipeak locations (having lower pairing probability and therefore likely to be more accessible) can be compared to the extendable sites detected by Allawi et al [39], which are indicated by green dots at the bottom of the plot. In each prediction, there are a number of apparently correct predictions and obvious errors, and it is not clear which method is yielding the most accurate results at the residue level. An experimental approach will eventually be required to determine which approach best represents the conditions of the microarray experiment, or if improvement is needed in both. In the absence of such validation, the Mfold accessibility predictions are sufficient to predict the scope of the secondary structure problem in a genome-based array design, even if some details of the prediction are not correct.

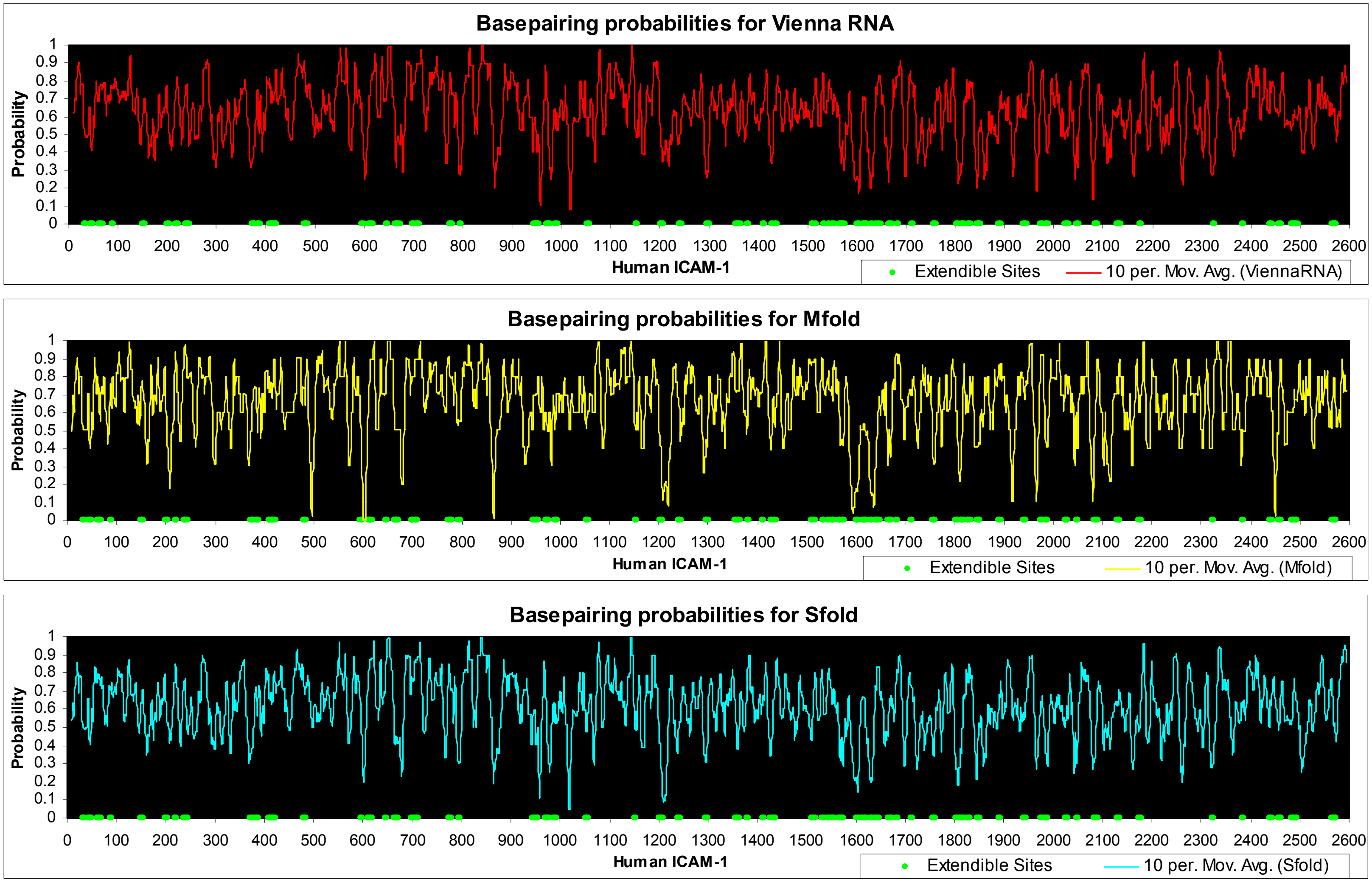


Figure 2.1.8.1 Accessibility prediction using three common methods. Pairing probabilities computed using RNAfold (top), Mfold (middle) and Sfold (bottom) for the human ICAM-1 transcript. Extendable sites detected by Allawi et al [39].

2.1.8.2 Loop length and other considerations

The focus of our study was specifically on the DNA/RNA base pairs that are actively involved in hydrogen bond formation. Other accessibility considerations will have to be taken into account in practice. The structure of a long single stranded DNA or RNA molecule can contain many nucleotides that, while not part of a double-helical stem, remain inaccessible to hybridization due to their location inside small loops, constrained by very stable hairpins at the ends, within the target secondary structure. A loop is a somewhat constrained structure as well, and the length at which it presents accessible sequence that favors hybridization has been shown to be on the order of 10 nucleotides and longer [36]. The nucleotides found in shorter loops may be classifiable as inaccessible. However, there is a need for quantitative hybridization experiments that would elucidate how loops and loop-like structures in long-oligo probe and target molecules affect the performance of assays.

Development of a target secondary structure criterion for oligonucleotide array design is expected to impose restrictions on the probe selection beyond the sequence similarity and melting temperature criteria that are currently used, especially in cases where short probe length restricts the annealing temperature used in the hybridization protocol to 22–37°. In the *B. suis* example, use of a low annealing temperature, e.g. 42°C which is the temperature used in some published 70-mer array experiments [3], would result in only about 30% of the average transcript being accessible for intermolecular hybridization, and this estimate does not include an accounting of the bases that are inaccessible by reason of being in short loops. Therefore, recommended hybridization temperatures for long synthetic oligomer arrays should be closer to 65°C, at which temperature our modeling of *Brucella* transcripts indicates that only 50% of a typical RNA transcript or 30% of the corresponding cDNA molecule remain inaccessible.

2.1.8.3 To shear or not to shear

While shearing reduces overall ΔG of secondary structure formation for individual molecules in the target solution, shearing does not in itself eliminate formation of secondary structure in single-stranded DNA or RNA. While some signal may be gained by reducing the stability of secondary structure in the target molecule, random shearing by its nature creates a mixture of targets that may have substantially different affinities. For instance, in a 300 nt transcript that is targeted by a 70mer oligonucleotide, there is nearly a one in four chance that a random break in the sequence will occur within the target site for which the probe is designed. Broken short fragments represent a different binding site with a different binding affinity than the full-length transcript. Binding of a 50mer sheared fragment to a 70mer probe leaves a dangling end in the probe. A break very close to one end or the other of the target site may create a target that still binds to the probe, though with reduced affinity; a break closer to the middle of the target site may produce fragments that bind partially to the probe, competing for binding with perfect matches. Thus the problem of modeling the components of the sample mixture becomes very much more difficult with a sheared sample.

2.1.8.4 The utility of experimentally validated biophysical criteria

The impact of secondary structure in single stranded polynucleotides on microarray results has been recognized and is being systematically studied [32-34, 39]. Intramolecular folding of mRNAs is so extensive that only 5–10% of most transcripts is accessible to binding of complementary nucleic acids; however the modeling of long molecules has not proven to give very accurate binding predictions [40]. Studies have demonstrated that, at 37°C and 0 mM $Mg^{2+}$and low concentration conditions, oligonucleotides that are more than 20 residues long yield better binding compared to shorter oligonucleotides [41]. Systematic "scanning" of mRNA sequences with libraries of short oligos [42] has also been shown to be successful in locating sites for siRNA targeting; however, such methods are likely to become extremely expensive if applied to the large number of targets in a microarray design.

2.1.9 Conclusion

A large fraction of the target molecule is predicted to be actively involved in formation of the stable secondary structure even at high temperature. Stable secondary structure in the target has the potential to interfere with hybridization and should be a factor in interpretation of microarray results, as well as an explicit criterion in array design. Ability to avoid the regions with the stable secondary structures on the target in an oligonucleotide design procedure would significantly change the definition of an optimal oligonucleotide.

The results of our study predict an important and overlooked role of target secondary structure in estimating the concentration of the target in a sample mixture when the assay is hybridization to oligonucleotide arrays. Oligonucleotide probe binding sites are found in double-stranded conformations in a significant fraction of transcripts even in cases where self-complementarity was avoided during the probe design process. At 52°C approximately 57% of probes designed for *Brucella* had binding sites in the target predicted to contain a stretch of unpaired bases of at least 14 nt in length; at 65°C, that fraction increased to 93%. Therefore, we would expect that at 52°C only 57% of all probes would demonstrate the expected behavior in the experiment. We predict that the remaining probes will exhibit modified binding behavior, and we plan to conduct experiments to characterize this behavior.

2.2 PROPERTIES OF *BRUCELLA* ARRAY

The second objective of the performed research was to investigate the molecular properties of *B. abortus, B. melitensis* and *B. suis* array.

The quality and specificity analysis of a set of 70mer probes designed by our collaborators at TIGR was assessed. These 70mer probes (3369 in total) target every predicted gene in the *B. suis* genome (NCBI, 11/03), and the arrays fabricated using these probes will allow measurement of gene transcription in *B. suis* under various conditions. Most of the probes on this array were designed based on the sequence of *B. suis*, using the program OligoArraySelector also known as pick70 [25]. An additional set of probes was designed based on the sequences of *B. abortus* and *B. melitensis*, to target genes which occur in these genomes but not in *B. suis*. The analysis included the calculation of several biophysical characteristics (such as the melting temperatures for the secondary structure formation and Gibbs free energies), and evaluation of the probe specificity. The specificity of these probes was investigated not only for transcripts from *B. suis*, but for sequences from *B. abortus* and *B. melitensis* as well. The sequences of the three genomes are sufficiently similar that transcripts from *B. abortus* and *B. melitensis* can be expected to hybridize to probes on the *B. suis* array. Below is the summary of the performed probe analysis.

The total number of the TIGR designed probes for comparative *Brucella* microarray is 3375. Seven ORFs on chromosome I and six ORFs on chromosome II of *B. suis* had no probes designed for them by TIGR.

2.2.1. Sequence Specificity

Since the majority of *Brucella* microarray probes were designed based on the *B. suis* genomic sequence and annotation, the specificity of these probes was checked against *B. abortus* and *B. melitensis* genomes. This analysis was performed using a standard BLASTn search having as parameters the expectation value of 10, word size 11 and the low complexity sequence filter.

The analysis revealed that 89% (3002) of the probes were unique within *Brucella*; however, the results produced multiple hits when the target database searched was the complete GenBank database (May 2003), which included all organisms.

Nine percent (305) of the TIGR designed probes were highly specific to *Brucella* and gave less than three partial hits in all other organisms. Fifty-five of these probe sequences were unique to *Brucella* and had no partial matches in any other prokaryotic or eukaryotic organism. Therefore, if transcription of the genes of these fifty-five probes is detected, these probes can be used as indicators of the representatives of genus *Brucella* on a diagnostic microarray. The rest of the 250 probes had one, two or three partial matches outside the *Brucella* genomes, and therefore, could be used as supporting *Brucella* identifiers.

This analysis also revealed a group of 68 low specificity probe sequences that were part of the *Brucella* array oligo set. Exactly one-half of these sequences had multiple (10 to 100) complete and partial matches to various open reading frames (ORFs) within the *Brucella* genomes, including two probes that matched about 100 different ORFs in *Brucella*, 16 probes that matched 50 ORFs and 4 probes with 30 to 40 hits in *Brucella* genomes. The gene descriptions for these ORFs included transposases and ribosomal RNAs. The rest of the low specificity probes included oligomers having less than 10 matches in the *Brucella* genomes. These probes may hybridize promiscuously to the transcripts of homologous genes, complicating the analysis of the microarray results. The last group of probes should be replaced with the different oligomers with the better sequence specificity, or should be used as a category of control probe.

2.2.2 Melting Temperatures

Oligomer probes can bind to themselves creating different kinds of loops and render parts of a sequence's sites inaccessible for target hybridization. The microarray probe-target hybridization step is usually performed at temperatures below 65˚C. Therefore, any probe structures stable at higher temperatures than this are undesired, and may affect the hybridization of these two nucleic acid molecules. Our analysis of the melting temperatures of the probes' secondary structures designed using the OligoArraySelector showed that only 807 oligomers had their T_ms below 65˚C. 2447 probes had their melting temperatures above 65˚C, of these 353 were above 80˚C and 55 were above 90˚C. This high secondary structure T_ms may create problems during the probe-target hybridization since it predicts a competing stable structure.

2.2.3 Polynucleotide Stretches

We also looked at continuous poly(Nt) stretches within the oligo probes. Affymetrix has published a study describing empirically derived and experimentally tested heuristics rules for designing short oligo probes [24]. According to Lockhart it is not recommended to have more than 5 Cs or Gs or more than 6 As or Ts in a row. These rules essentially enforce a lower bound of complexity, or minimum information content to prevent nonspecific probe binding. In the set of *Brucella* probes used, the following data for the continuous stretches of 5-7 nucleotides, were noted and may be present twice within one sequence:
polyG – 76 probes,
polyC – 47 probes,
polyA – 291 probes
polyT – 301 probes.

2.2.4 Mismatches

The possible mis-alignment of the *B. suis* designed oligo probes when used for the microarray analysis of *B. melitensis* was investigated. Our analysis shows that in a number of cases the probes are complementary to the genomic DNA sequence, but are located outside the ORF borders. Such irregular probe locations together with the experimental microarray data will allow a re-evaluation of the *B. melitensis* and *B. abortus* annotations. The probes for the mentioned ORFs should be re-designed to match all three genomes based on the existing annotations.

Currently the biophysical data collection for the *Brucella* microarray probes is accumulated in a very large Excel spreadsheet. The spreadsheet contains the information on location of the *Brucella* ORF on the chromosome down to the nucleotide number, length of the target mRNA molecule, the gene name and function or hypothetical ORF function, oligo ID and sequence, polyNt stretches information, ΔG, ΔH, ΔS and secondary structure Tm values as well as the information on missannealing of the probes with regards to *B. melitensis* and *B. abortus* sequence annotation and genomic DNA differences. A few useful additions would include addition of the probe G/C percentage and probe misplacement drawings for *B. abortus*, as well as the analysis of target stable secondary structures. It would be really useful to link data from the spreadsheet to the microarray layout (or Array Design in MAGE terms) map, so that by clicking on a spot within an image one could focus on the nature of the probe placed on the microarray chip.

2.3. IDENTIFICATION OF FEATURES WITH DIAGNOSTIC UTILITY IN *BRUCELLA*

The third objective of the proposed research is to identify nucleotide stretches with diagnostic utility located in the differential regions of the three Brucella *species.*

2.3.1 RT-PCR analysis of differential regions in *Brucella*

The three way comparison of the *Brucella* spp.: *B. abortus*, *B. melitensis* and *B. suis* revealed a total of 23 unique genes, and 79 differential genes common between any two of the three species. The reverse transcriptase (RT-PCR) analyses of these differentiating regions revealed that several groups of genes with potential significance for virulence are transcribed *in vivo*. Some of these genes are likely to be of phage or plasmid origin, suggesting possible mechanisms for their appearance as differentials [43]. Table 2.3.1 summarizes transcripts detected for genes in each differentiating sequence island.

Table 2.3.1 Summary of RT-PCR results for differential ORFs.

Genomic	***Brucella suis***			***Brucella melitensis***			***Brucella abortus***		
Location	**Predicted**	**Observed**	**NB**	**Predicted**	**Observed**	**NB**	**Predicted**	**Observed**	**NB**
S1	4	4	0	1	1	0	0	0	0
S2	18	17	1	0	0	0	0	0	0
M1	0	0	0	1	1	0	0	0	0
A	0	0	0	0	0	0	1	1	0
SM1	1	1	0	2	2	0	0	0	0
SM2	25	16	9	24	6	18	0	0	0
SA1	11	3	8	0	0	0	9	7	2
SA2	11	7	4	0	0	0	11	6	5
MA1	0	0	0	30	21	9	26	23	3

A – *B. abortus*, M – *B. melitensis*, S – *B. suis*, 1 – chromosome I and 2 – chromosome 2 and NB – no band.

Possible uses for differential targets include:

- define probes for a diagnostic miniarray,

- clarify and validate the RT-PCR results obtained earlier for the differential regions,
- test the transcription of the differential regions in *Brucella*, using the probes designed for the common and differentiating parts of the ORFs, which allows several probes to be placed on one differential ORF. For example, a differentiating ORF from the MA1 island has a long region common to the both ORFs from *Brucella melitensis* and *B. abortus* chromosome I, and another smaller region unique to *B. melitensis.* In this case there are two probes for the differential ORFs: one located in the common part of the two genes and the other one located in the *B. melitensis* unique part of the gene. Generalizing this approach should help clarify whether the differences in the ORFs' length and sequence are due to annotation errors or reflect the real evolutionary divergence between the *Brucella,*
- compare the quality of the oligo probes designed using two different types of software, whose methods either incorporate or ignore a step that verifies probe secondary structure formation,
- and test the transcription of the fifty-five probes from the ORFs shared by all three Brucella species predicted to be unique to *Brucella* and not any other prokaryotic or eukaryotic organism with the known genome.

2.3.2 PCR Analysis of 18 Different *Brucella* Biovars

The whole genome comparison of three *Brucella* species (*B. suis 1330, B. melitensis 16M* and *B. abortus 544*) and identification of unique and differential regions in these pathogenic bacteria provided the basis for the identification of *Brucella* spp. The knowledge of the complete genome sequences from three *Brucella* biovars, representing three different *Brucella* species and the regions that differ between these genomes can be used to set up the PCR reactions, which will help to discriminate between all six *Brucella* species and may be even some of its biovars. Since, our knowledge of the complete *Brucella* genome sequence is limited to only three out of over twenty different biovars, there is a chance that a PCR primer pair designed to amplify a unique region in particular biovars with known genome may produce a band in different *Brucella* species, whose genome has not been sequenced yet. There is also a chance to have no PCR product or a PCR product of a different size, when amplifying genomic DNA in the other (not sequenced) biovars of *B. suis, B. melitensis* and *B. abortus.*

In order to test the applicability of the unique and differentiating regions for *Brucella* diagnostics 24 out of more than 100 computationally predicted and experimentally validated open reading frames were chosen with the intent of finding a combination that would identify each of the three *Brucella* species. The PCR was performed on methanol killed bacterial cells from eighteen different *Brucella* biovars, which is a commonly used diagnostic technique [44, 45]. The eighteen tested *Brucella* biovars included the representatives of all six species of *Brucella.* These included 5 different biovars of *B. suis* (*B. suis 1330* biovar 1, *B. suis Thompsen* biovar 2, *B. suis 686* biovar 3, *B. suis 40* biovar 4 and *B. suis 513* biovar 5) *,* three biovars of *B. melitensis* (*B. melitensis 16M* biovar 1, *B. melitensis 63/9* biovar 2 and *B. melitensis Ether* biovar 3), seven representatives of *B. abortus* (*B. abortus 544* biovar 1, *B. abortus 86/8/59* biovar 2, *B. abortus Tulya* biovar 3, *B. abortus 292* biovar 4, *B. abortus B3196* biovar 5, *B. abortus 870* biovar 6 and *B. abortus C68* biovar 9), *B. ovis 1155*, *B. neotomae 5K33* and *B. canis RM 6/66*.

The cells obtained from Dr. Betsy Bricker were stored in 66% methanol and the night before the PCR analysis diluted in water down to 0.2-0.15 OD_{550} nm. PCR was performed using the Invitrogen PCR Supermix, a 55°C annealing temperature and 1 min elongation time. The PCR products then were separated on the 1.6% agarose gel in a sodium borate buffer.

Ten of the PCR reactions worked out unambiguously (meaning they produced specific single bands of expected size when predicted computationally and no PCR amplification was observed in the biovars, where these sequence fragments were expected to be absent based on their known genomic sequences). A *Brucella* genus-specific PCR primer pair was used as a positive control for the genomic DNA from the methanol killed bacterial cells. This control primer pair was screened against all known sequences from all organisms (including the all of the eukaryotes) currently entered in the GenBank Database, and based on these results is expected to be extremely *Brucella* specific. Each set of PCR reactions contained no DNA contamination control. The primer pair sequences and expected PCR product sizes are shown in the Table 2.3.2.a.

Table 2.3.2.a Primer pairs used for PCR amplification of unique and differential regions in 18 *Brucella* biovars.

Primer Pair number	Forward primer	Reverse primer	ORF name and Amplicon Size, bp			Gene function
			B. abortus	*B. melitensis*	*B. suis*	
1	TGATAGCGCCAGACAACAAC	TGTGCCAGCTTCGTTGTAAG	BruAb1_1825 596	BMEI0205 470	BR1846 722	Immunoglobulin-binding protein EIBE
2	AAATGTCAATCTGGGCTTCG	TATTGAAGAACTGCGCAACG			BRA0378 191	Hypothetical protein
3	ATTTATGTCCGTGAACTGTCCGTC	TTGTCCGCAAAAAGTATCAAAACG			BRA0369 123	Hypothetical protein
4	AACTGCTGGAGATGAATCCG	GAATGTTTGCACGCATCAAT			BRA0363 149	DNA-binding protein
5	CTTTACGCCCGTGTATCGAC	CATGGGGTCCTGTGTTGAG		BMEI1661 321		Recombinase
6	TGCAGCTCACGGATAATTTG	ACACCTTGTCCACGCTCAC	BruAb2_0168 783			Outermembrane transporter
7	AGCTTCTGGAGGAGGTGGAT	GTTCCGCCTTGTGTTTCTTC		BMEII0827 526	BRA0439 526	Glucose-1-phosphate cytidylyltransferase
8	TCTACACCACGCTGAAGTCG	CCGAAAGCCGATAGAGTTTG	BruAb2_1035 393	BMEII0204 162	BRA1096 393	Transcriptional regulator, GNTR family
9	TTGTTGGAAACGGCTTTGATATC	GAAAGTACCCACCCTCGGAAAACT	BruAb1_0266 358	BMEI1681 358		Hypothetical protein
10	TCATGCTGTGCCTCCAATTCC	TTGCTGAGCAGCAGCAAGAAC	BruAb1_0248 184	BMEI1699 184		Hypothetical protein
Control	TCAGGCGCTTATAACCGAAG	ATCTGCGCATAGGTCTGCTT	BruAb2_0582 261	BMEII0637 261	BRA0644 261	pcaC 4-carboxymuconolactone decarboxylase

The PCR results we obtained with these primer pairs agreed with the computational predictions, and are summarized in Table 2.3.2.a. These results indicate that the computationally predicted unique and differential *Brucella* regions can be used to build the diagnostic tests for the known *Brucella* species and some biovars. For example, the PCR amplification of a small portion of a 6 kb long coding sequence from *B. abortus* produces specific band in every one of the seven different tested *B. abortus* biovars, and not in any other *Brucella* species. Therefore, although there is no unique open reading frame predicted for *Brucella abortus*, there is at least 783 base long region in this species that is present in *B. abortus* only, and is not found in the other *Brucella.* The *B. abortus* primer pair was also screened against the entire GenBank Database and shows extremely high specificity. Another example of possible diagnostic application of PCR amplification in is the primer pair number 8 in the Table 2.3.2.b. These PCR primer pair was originally designed based on the sequence of one of the open reading frames present only in *B. suis* and *B. abortus*, but turned out to also give same size fragments in *B. neotomae* and *B. canis*, and give a substantially shorter size fragment in all three tested *B. melitensis* biovars. This primer pair did not produce a fragment in *B. ovis* biovar. Since the primer pair designed to amplify a single unique *B. melitensis* ORF has successfully amplified the regions in several other *Brucella* species (including *B. ovis*, *B. neotomae* and *B. suis*) and did not produce an amplicon in *B. melitensis 63/9*, it can only be used as a diagnostic helper, but not as a unique identifying feature. In the light of the above information, the primer pair number 8 becomes especially important, as a tool capable to differentiate *B. melitensis* biovars from the other *Brucella* species. It is also important to mention that although, computational analysis pointed out the presence of large differentiating islands present in *Brucella suis 1330* genome, the PCR reactions performed on several ORFs located in these islands indicate that, while similar ORFs are present in *B. canis RM* and *B. neotomae 5K33*, *B. suis 513* seems to be missing all of the tested ORFs! This indicates that *B. canis*, *B. neotomae* and *B. suis* may be related evolutionary. From the evolutionary point of view it is interesting to mention that *Brucella suis Thompsen* produced the largest number of bands when tested with all 24 unique and differential primer pairs (data not shown), and therefore seems to have the most ORF rich genome than any other of the eighteen investigated *Brucella* biovars. Further investigation of all one hundred unique and differential regions may reveal the presence of other sites in *Brucella* genomes capable to unambiguously differentiate certain species and perhaps even their biovars. In addition, the extremely high *Brucella* specificity of another two primer pairs # 5 and #9 should be noted due to its potential for diagnostic application.

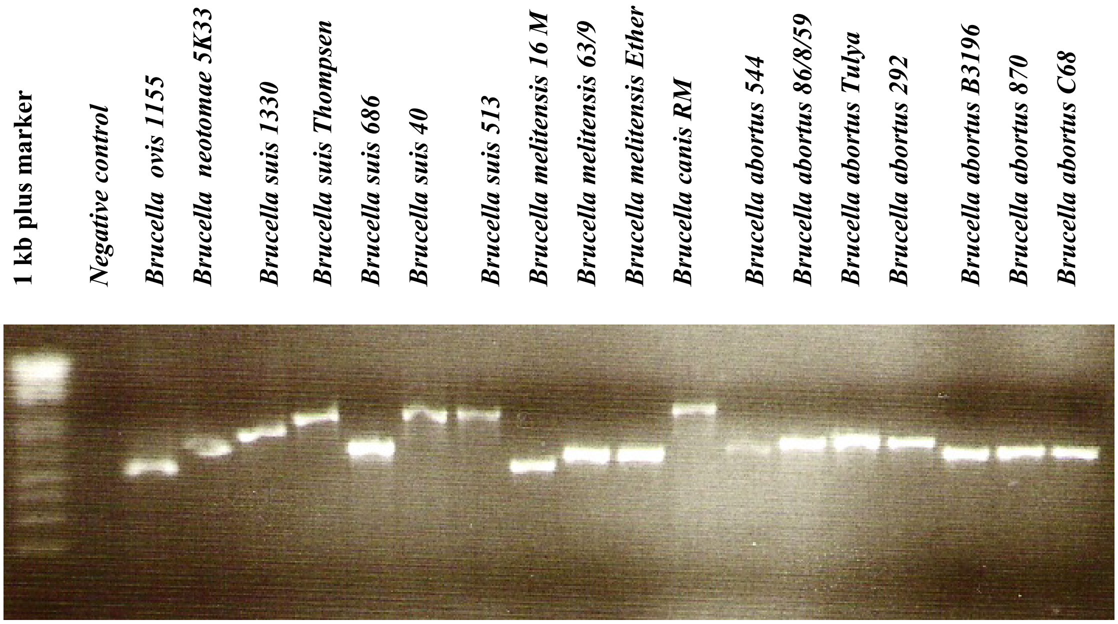
Table 2.3.2.b Summary of the PCR screening of 18 *Brucella* biovars.

Primer pair	*Brucella ovis 1155*	*Brucella neotomae 5K33*	*Brucella suis 1330*	*Brucella suis Thompsen*	*Brucella suis 686*	*Brucella suis 40*	*Brucella suis 513*	*Brucella melitensis 16 M*	*Brucella melitensis 63/9*	*Brucella melitensis Ether*	*Brucella canis RM*	*Brucella abortus 544*	*Brucella abortus 86/8/59*	*Brucella abortus Tulya*	*Brucella abortus 292*	*Brucella abortus B3196*	*Brucella abortus 870*	*Brucella abortus C68*
1	🟩	🟩	▬	🟩	🟩	🟩	🟩	▬	🟩	🟩	🟩	▬	🟩	🟩	🟩	🟩	🟩	🟩
2		▬	▬	▬	▬	▬					▬							
3		▬	▬	▬	▬	▬					▬							
4		▬	▬	▬	▬	▬					▬							
5	▬	▬		▬	–		▬	▬		▬								
6												▬	▬	▬	▬	▬	▬	▬
7	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬		–	–	–	–	–	–
8		▬	▬	▬	▬	▬	▬	🟩	🟩	🟩	▬	▬	▬	▬	▬	▬	▬	▬
9	▬	▬		▬				▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	
10	▬	▬		▬			▬	▬	▬	▬		▬	▬	▬	▬	▬	▬	▬
Control	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬

Small dashes represent weak bands of the correct size observed on the gel. Green bands show size variability.

Primer pair #1 was designed for PCR amplification of a fragment within a unique *B. suis 1330* ORF. Based on the computational analysis we also expected that this primer pair will amplify a 722 bp. long fragment in *B. suis 1330*, as well as two shorter fragments of different length in *B. melitensis 16M* and *B. abortus 544.* This prediction proved to be true, as seen in the Figure 2.3.2.

Figure 2.3.2 PCR amplification of a polymorphic region in *Brucella.*



In addition to obtaining the bands of expected size in these three biovars, we also obtained a band in every one of the other 15 *Brucella* biovars. The size of the PCR product produced varied between the species and even biovars. Sequencing of every band amplified with this primer pairs will help to further investigate this polymorphic region.

3. Future Work

3.1 Target Secondary Structure Modeling

The results obtained in the study, which involved the modeling of the target secondary structure, strongly suggest performing the following investigations:

- use a software other than Mfold that calculates the ensemble weighted free energy change to refine and validate the overall picture of the target molecules occupied by the secondary structure during their on chip hybridization;
- study experimentally the effects of target secondary structure formation on hybridization to the corresponding probe; design and perform a miniarray experiment specifically to test the effects of probe placement relative to location of stable secondary structure, using either *Brucella* genes or genes from a model organism. For this purpose a new mini-array, will be designed to contain several probe-target pairs array with the secondary structure formation on the target molecules predicted to interfere with probe hybridization. Also design and add probes for the same target molecules that will anneal to regions predicted to consist of regions of random coil. Model the sequences with the known hybridizational behavior can be used as our positive controls. The probe-target pairs with no stable secondary structure formations can be used to verify the particular gene expression. We will utilize available software to calculate the hybridization kinetics. Using artificial target mixtures of known concentration we will compare the signal intensity between the probe-target pairs with predicted secondary structure and those where it is predicted or known to be absent. This mini-array will provide experimental support to our modeling investigations of the effect of the target secondary structure on the probe-target duplex formation;
- investigate the role of the target bases hidden inside of the different loops and 3D structures on the target molecule's ability to bind to its probe;
- develop the target stable secondary structure into a criterion, for the microarray probe design software;
- and develop an automated procedure for annotating the targets for the regions that are likely to be inaccessible under the hybridization conditions. This software will search for the

accessible regions on the target molecules suitable for the probe placement, and will generate the biophysical table for the properties of the target molecules.

3.2 PCR FINGEPRINTING FOR *BRUCELLA*

Further investigations of the differential regions in *Brucella* may involve PCR and RT-PCR screening of all identified regions across a number of known biovars. This will elucidate functional importance of the observed nucleotide differences among the *Brucella* species, and will help to create novel diagnostic tests for *Brucella*, as well as will clarify the evolutionary relationship between these extremely closely related species.

Molecular weight	***Brucella abortus***	***Brucella melitensis***	***Brucella ovis***	***Brucella canis, suis* and *neotomae***
162		—		
261	—	—	—	—
393	—			—
783	—			

Table 3.2. *Brucella* PCR fingerprinting.

Based on the experimental results obtained in this study it should be possible to develop a PCR fingerprint for each known *Brucella* species using multiple PCR primer pairs in one reaction. Since the purpose of this investigation was to validate the usefulness of computationally predicted differential regions for the diagnostic purposes rather than development of a complete diagnostic test, thoroughly investigate all the 100 differential regions and primer pairs were no thoroughly investigated. However, results produced here are sufficient to identify three *Brucella* species purely based on the genomic PCR. Table 3.2 shows a schematic gel representation of the PCR results obtained in this study using the differential primer pairs #6, #8 and a primer pair #11 unique to *Brucella* genus.

4. References Cited

1. Shi, S.J., et al., *DNA exhibits multi-stranded binding recognition on glass microarrays.* Nucleic Acids Res, 2001. **29**(20): p. 4251-6.
2. Yue, H., et al., *An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression.* Nucleic Acids Res, 2001. **29**(8): p. E41-1.
3. Wang, H.Y., et al., *Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays.* Genome Biol, 2003. **4**(1): p. R5.
4. Kane, M.D., et al., *Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays.* Nucleic Acids Res, 2000. **28**(22): p. 4552-7.
5. Hughes, T.R., et al., *Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.* Nat Biotechnol, 2001. **19**(4): p. 342-7.
6. Ramakrishnan, R., et al., *An assessment of Motorola CodeLink microarray performance for gene expression profiling applications.* Nucleic Acids Res, 2002. **30**(7): p. e30.
7. Relogio, A., et al., *Optimization of oligonucleotide-based DNA microarrays.* Nucleic Acids Res, 2002. **30**(11): p. e51.
8. Zuker, M., *Calculating nucleic acid secondary structure.* Curr Opin Struct Biol, 2000. **10**(3): p. 303-10.
9. Mathews, D.H., et al., *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.* J Mol Biol, 1999. **288**(5): p. 911-40.
10. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
11. Rouillard, J.M., M. Zuker, and E. Gulari, *OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach.* Nucleic Acids Res, 2003. **31**(12): p. 3057-62.
12. Chou, C.C., et al., *Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression.* Nucleic Acids Res, 2004. **32**(12): p. e99.
13. Bommarito, S., N. Peyret, and J. SantaLucia, Jr., *Thermodynamic parameters for DNA sequences with dangling ends.* Nucleic Acids Res, 2000. **28**(9): p. 1929-34.
14. SantaLucia, J., Jr., *A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics.* Proc Natl Acad Sci U S A, 1998. **95**(4): p. 1460-5.
15. Chan, V., D.J. Graves, and S.E. McKenzie, *The biophysics of DNA hybridization with immobilized oligonucleotide probes.* Biophys J, 1995. **69**(6): p. 2243-55.
16. Gingeras, T.R., D.Y. Kwoh, and G.R. Davis, *Hybridization properties of immobilized nucleic acids.* Nucleic Acids Res, 1987. **15**(13): p. 5373-90.
17. Boschiroli, M.L., V. Foulongne, and D. O'Callaghan, *Brucellosis: a worldwide zoonosis.* Curr Opin Microbiol, 2001. **4**(1): p. 58-64.
18. Moreno, E., et al., *Brucella abortus 16S rRNA and lipid A reveal a phylogenetic relationship with members of the alpha-2 subdivision of the class Proteobacteria.* J Bacteriol, 1990. **172**(7): p. 3569-76.
19. Young, E.J., *Human brucellosis.* Rev Infect Dis, 1983. **5**(5): p. 821-42.
20. Halling, S.M., et al., *Completion of the genome sequence of Brucella abortus and comparison to the highly similar genomes of Brucella melitensis and Brucella suis.* J Bacteriol, 2005. **187**(8): p. 2715-26.
21. DelVecchio, V.G., et al., *The genome sequence of the facultative intracellular pathogen Brucella melitensis.* Proc Natl Acad Sci U S A, 2002. **99**(1): p. 443-8.

22. Paulsen, I.T., et al., *The Brucella suis genome reveals fundamental similarities between animal and plant pathogens and symbionts.* Proc Natl Acad Sci U S A, 2002. **99**(20): p. 13148-53.

23. Ratushna, V.G., J.W. Weller, and C.J. Gibas, *Secondary structure in the target as a confounding factor in synthetic oligomer microarray design.* BMC Genomics, 2005. **6**(1): p. 31.

24. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.

25. Bozdech, Z., et al., *Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray.* Genome Biol, 2003. **4**(2): p. R9.

26. Chou, H.H., et al., *Picky: oligo microarray design for large genomes.* Bioinformatics, 2004. **20**(17): p. 2893-902.

27. Nielsen, H.B., R. Wernersson, and S. Knudsen, *Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays.* Nucleic Acids Res, 2003. **31**(13): p. 3491-6.

28. Tolstrup, N., et al., *OligoDesign: optimal design of LNA (locked nucleic acid) oligonucleotide capture probes for gene expression profiling.* Nucl. Acids. Res., 2003. **31**(13): p. 3758-3762.

29. Nguyen, H.-K. and E.M. Southern, *Minimising the secondary structure of DNA targets by incorporation of a modified deoxynucleoside: implications for nucleic acid analysis by hybridization.* Nucl. Acids. Res., 2000. **28**(20): p. 3904-3909.

30. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406-15.

31. Ding, Y. and C.E. Lawrence, *Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond.* Nucleic Acids Res, 2001. **29**(5): p. 1034-46.

32. Ding, Y. and C.E. Lawrence, *A statistical sampling algorithm for RNA secondary structure prediction.* Nucleic Acids Res, 2003. **31**(24): p. 7280-301.

33. McCaskill, J.S., *The equilibrium partition function and base pair binding probabilities for RNA secondary structure.* Biopolymers, 1990. **29**(6-7): p. 1105-19.

34. Hofacker, *Fast folding and comparison of RNA secondary structures.* Monatshefte f Chemie, 1994.

35. Amarzguioui, M., et al., *Secondary structure prediction and in vitro accessibility of mRNA as tools in the selection of target sites for ribozymes.* Nucleic Acids Res, 2000. **28**(21): p. 4113-24.

36. Scherr, M., et al., *RNA accessibility prediction: a theoretical approach is consistent with experimental studies in cell extracts.* Nucleic Acids Res, 2000. **28**(13): p. 2455-61.

37. Kretschmer-Kazemi Far R, S.G., *The activity of siRNA in mammalian cells is related to structural target accessibility: a comparision with antisense oligonucleotides.* Nucleic Acids Res, 2003. **31**(15): p. 4417-4424.

38. Ding, Y., C.Y. Chan, and C.E. Lawrence, *Sfold web server for statistical folding and rational design of nucleic acids.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W135-41.

39. Allawi, H.T., et al., *Mapping of RNA accessible sites by extension of random oligonucleotide libraries with reverse transcriptase.* Rna, 2001. **7**(2): p. 314-27.

40. Sohail, M., S. Akhtar, and E.M. Southern, *The folding of large RNAs studied by hybridization to arrays of complementary oligonucleotides.* Rna, 1999. **5**(5): p. 646-55.

41. Bohula, E.A., et al., *The efficacy of small interfering RNAs targeted to the type 1 insulin-like growth factor receptor (IGF1R) is influenced by secondary structure in the IGF1R transcript.* J Biol Chem, 2003. **278**(18): p. 15991-7.

42. Zhang, H.Y., et al., *mRNA accessible site tagging (MAST): a novel high throughput method for selecting effective antisense oligonucleotides.* Nucleic Acids Res, 2003. **31**(14): p. e72.

43. Sturgill, D., *Comparative Genome Analysis of Three Brucella spp. and a Data Model for Automated Multiple Genome Comparison*, in *Biology*. 2003, Virginia Tech: Blacksburg, VA. p. 55.

44. Bricker, B.J., *PCR as a diagnostic tool for brucellosis.* Vet Microbiol, 2002. **90**(1-4): p. 435-46.

45. Bricker, B.J., *Diagnostic strategies used for the identification of Brucella.* Vet Microbiol, 2002. **90**(1-4): p. 433-4.

5.0 APPENDIX

BMC Genomics

Research article

Secondary structure in the target as a confounding factor in synthetic oligomer microarray design

Vladyslava G Ratushna[1], Jennifer W Weller[2] and Cynthia J Gibas*[1]

Address: [1]Department of Biology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 24061, USA and [2]School of Computational Science, Prince William Campus of George Mason University, Manassas, Virginia, 20110, USA

Email: Vladyslava G Ratushna - vratushn@vt.edu; Jennifer W Weller - jweller@gmu.edu; Cynthia J Gibas* - cgibas@vt.edu

* Corresponding author

Published: 08 March 2005

BMC Genomics 2005, **6**:31 doi:10.1186/1471-2164-6-31

Received: 10 September 2004

Accepted: 08 March 2005

This article is available from: http://www.biomedcentral.com/1471-2164/6/31

Abstract

Background: Secondary structure in the target is a property not usually considered in software applications for design of optimal custom oligonucleotide probes. It is frequently assumed that eliminating self-complementarity, or screening for secondary structure in the probe, is sufficient to avoid interference with hybridization by stable secondary structures in the probe binding site. Prediction and thermodynamic analysis of secondary structure formation in a genome-wide set of transcripts from *Brucella suis 1330* demonstrates that the properties of the target molecule have the potential to strongly influence the rate and extent of hybridization between transcript and tethered oligonucleotide probe in a microarray experiment.

Results: Despite the relatively high hybridization temperatures and 1M monovalent salt imposed in the modeling process to approximate hybridization conditions used in the laboratory, we find that parts of the target molecules are likely to be inaccessible to intermolecular hybridization due to the formation of stable intramolecular secondary structure. For example, at 65°C, 28 ± 7% of the average cDNA target sequence is predicted to be inaccessible to hybridization. We also analyzed the specific binding sites of a set of 70mer probes previously designed for *Brucella* using a freely available oligo design software package. 21 ± 13% of the nucleotides in each probe binding site are within a double-stranded structure in over half of the folds predicted for the cDNA target at 65°C. The intramolecular structures formed are more stable and extensive when an RNA target is modeled rather than cDNA. When random shearing of the target is modeled for fragments of 200, 100 and 50 nt, an overall destabilization of secondary structure is predicted, but shearing does not eliminate secondary structure.

Conclusion: Secondary structure in the target is pervasive, and a significant fraction of the target is found in double stranded conformations even at high temperature. Stable structure in the target has the potential to interfere with hybridization and should be a factor in interpretation of microarray results, as well as an explicit criterion in array design. Inclusion of this property in an oligonucleotide design procedure would change the definition of an optimal oligonucleotide significantly.

Background

Sequence-specific hybridization of a long single-stranded labeled DNA or RNA target molecule to shorter oligonucleotide probes is the basis of the gene expression microarray experiment. In this type of microarray experiment, gene specific *probe* molecules are either synthesized in situ or are printed to the microarray slide, and are either nonspecifically cross-linked to the surface or are attached specifically using a method such as poly-Lysine linkers. *Target* molecules (most often fluorescently labeled cDNA molecules, although cRNA and aRNA are used in some protocols) hybridize transiently to the probe oligomers until they form stable double helices with their specific probes. At some point, the rate of on and off reactions reach equilibrium, and the concentration of the target in the sample solution can be calculated. Transcript abundance is assessed by the relative intensity of signal from each spot on the array. This interpretation of array data relies on the assumption that each hybridization reaction goes to completion within the timeframe of the experiment and that the behavior of all pairs of intended reaction partners in the experiment is somewhat uniform.

There are three major types of DNA microarrays, which differ in the approach used for probe design: Affymetrix type microarrays [1], which assay each transcript with a distributed set of 25-mer oligonucleotides, full length cDNA microarrays, in which long cDNA molecules of lengths up to several hundred bases are crosslinked to the slide surface to probe their complement [2], and synthetic long-oligomer probe microarrays, which usually assay each transcript only once. The latter class of microarrays encompasses a variety of commercial and custom platforms, and there has yet to emerge a consensus on an optimal probe length for particular experimental designs. Oligo lengths ranging from 35 to 70 nucleotides have been shown to perform well under different conditions [3-5], though recent studies have shown that oligomers of up to 150 nucleotides may be desirable for assessing transcript abundance [6]. In general, the use of synthetic oligomers has been shown to result in improved data quality [7,8] relative to cDNA arrays, and 70mers have been shown to detect target with a sensitivity similar to that of full length cDNA probes [9]. Short probes have been promoted because they facilitate finding unique sequence matches while forming fewer, and less stable, hairpin structures and because they display more uniform hybridization behavior overall. However, the need for sensitivity and detection of transcripts in low copy number drives the use of long-oligonucleotide arrays. In this study, we have modeled the accessibility of transcripts to hybridization with 70mer oligonucleotides.

A number of oligonucleotide design software packages have been published in recent years, each having design strengths in one of a number of criteria [10-14]. Several factors are considered by almost all microarray design software packages: in particular, the sequence specificity of the probe-target interface and the overall balance of GC content across the array. Unique regions of the target sequence are identified using sequence comparison methods; the unique regions become the search space for probe selection based on other criteria. The number of probes per sequence and location of the probe in the sequence also restrict sequence availability. A relatively uniform melting profile generally is achieved simply by selecting for probes with similar GC content and uniform or close-to-uniform length, although some design methods explicitly compute the duplex melting temperature for each candidate probe-target pair and filter unique probes to find those which match a specified range of melting temperatures. Another biophysical criterion that is sometimes applied is the elimination of probes having the ability to form stable intramolecular structures under the conditions of the experiment. This is usually done by eliminating regions of self-complementarity, although at least one design program [13] does explicitly compute the melting temperature of the most stable structure to form in the probe molecule and uses that information to filter out stable secondary structures in the probe.

Few of the available array design packages explicitly consider the possible structures of the transcript-derived molecules in the sample solution and their impact on whether the microarray will provide an effective assay, although the OligoDesign web server [14] does compute this information for use in design of locked nucleic acid probes. It has been shown that a hairpin of as little as six bases in an oligonucleotide can require a 600-fold excess of the complementary strand to displace the hairpin even partially [15]. Since the target molecules are generally longer than the probe and may be of a different chemistry, it is not sufficient to conclude that their behavior will mirror that of the complementary probe. Prediction of secondary structure in a sample transcript using a standard nucleic acid secondary structure prediction algorithm (Mfold) demonstrates that while longer-range interactions are reduced at high temperatures, stable local structures persist in the transcript even at high salt concentration and high temperature (Figure 1). Because unimolecular reactions within the target can occur on a much shorter timescale than the diffusion-mediated, bimolecular, duplex hybridization reaction, competition for binding by intramolecular structures is expected to block the specific probe annealing sites on the target sequence in some cases and result in misinterpretation of the signal obtained from the assay if these effects are not taken into account.

In order to estimate the prevalence of stable secondary structure in long target molecules, and thus the impact

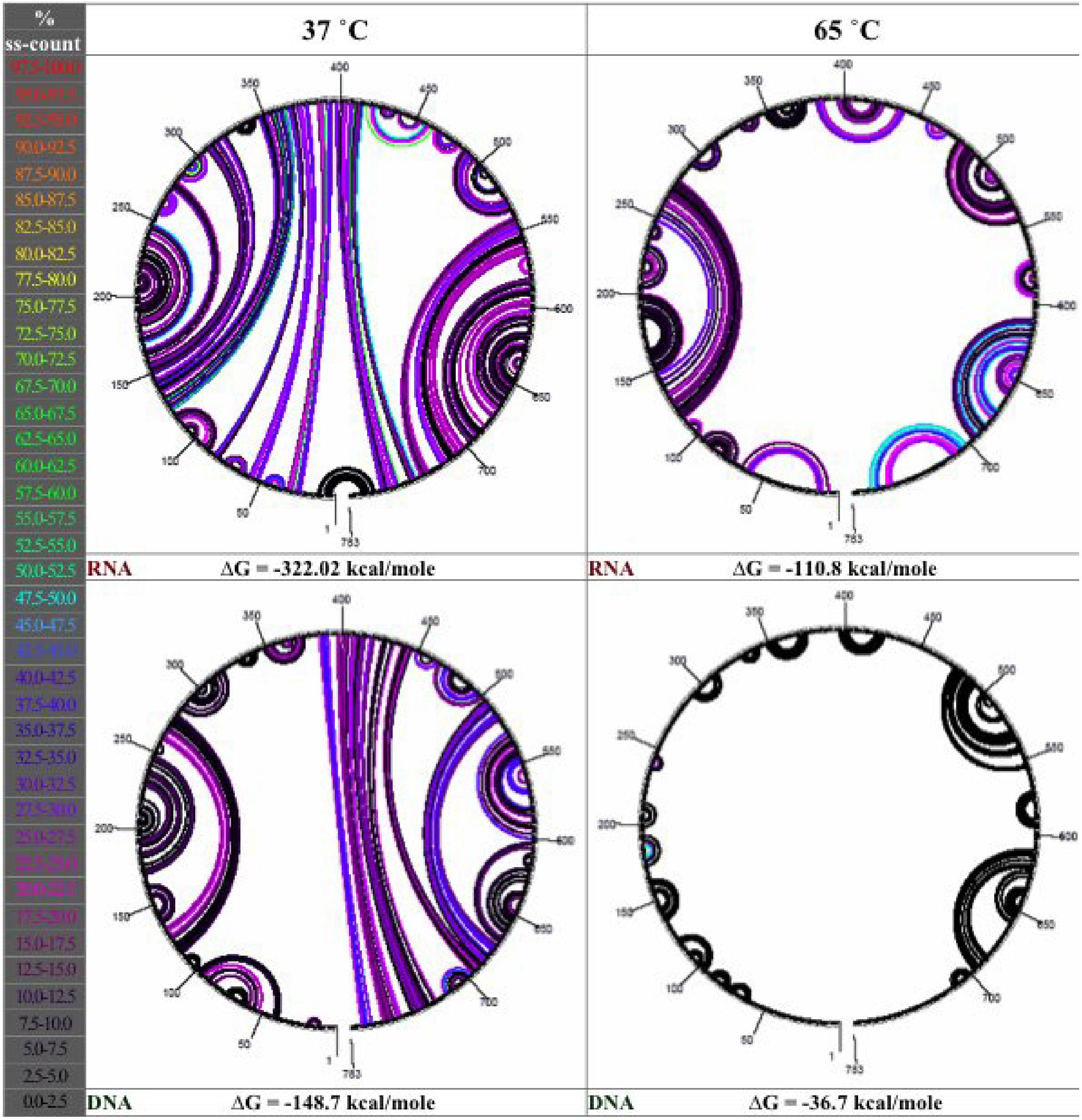


Figure 1

Secondary structure in a sample transcript. Circular diagrams of structure in a sample transcript (moeB homolog designated BR0004) from *Brucella suis*. Circular diagrams show hydrogen bonds between individual nucleotides, color-coded according to single-strandedness – the fraction of structures in which that bond is not present. Black bonds indicate 0% single-strandedness; red bonds indicate 100% single-strandedness.

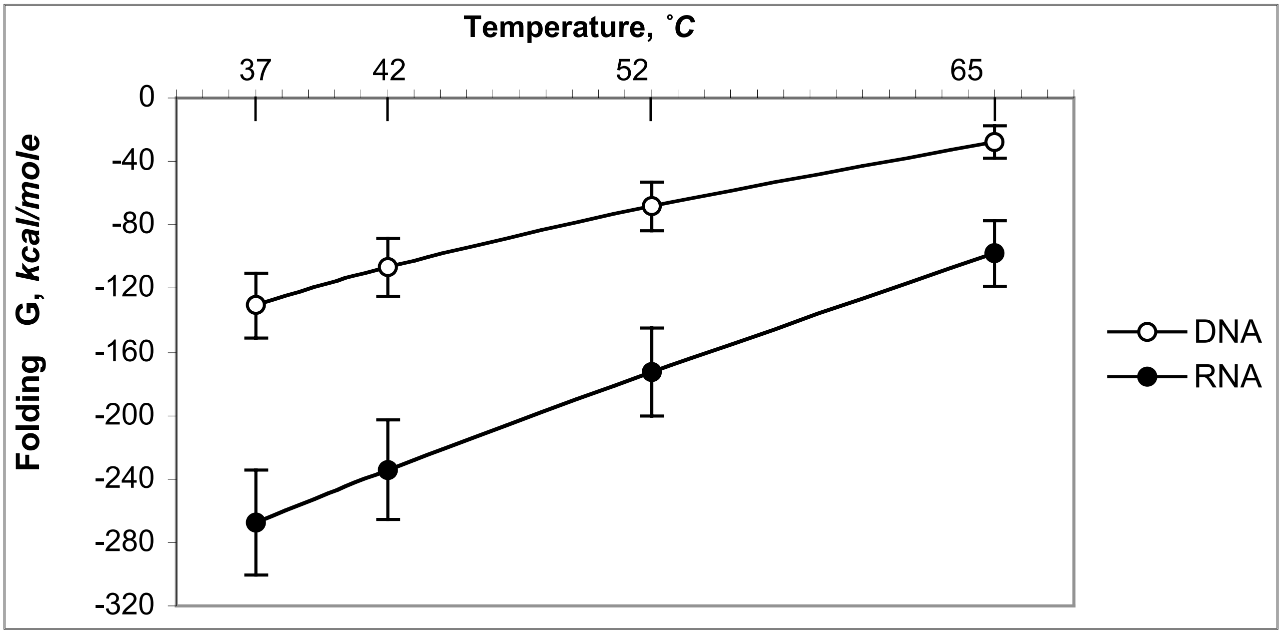


Figure 2
Stability of transcript secondary structure in *Brucella suis.* Average free energy change on global secondary structure formation for *Brucella suis* targets, modeled as DNA or RNA. ΔG values are normalized to global mean target length.

such structures might have on the analysis of microarray data, we have modeled secondary structure formation in mRNA transcripts of the intracellular pathogen *Brucella suis*. We have assessed the stability of structures formed in the transcript and the accessibility of the binding sites of optimal probes generated using commonly applied design criteria. Because random shearing of the full-length target molecule is used in some protocols, we have also modeled the effects of shearing to an average length on the prevalence of secondary structure in selected targets.

Results

Extent and stability of target secondary structure

Our modeling results obtained for the genome-wide set of intact single-stranded DNA or RNA targets demonstrate that stable secondary structures are widespread in target mixtures from *Brucella suis* (Figure 2) and in randomly chosen transcripts from the genomes of *E. coli* and *L. lactis*. Figure 2 shows the ΔG of formation for the most stable predicted secondary structure of the full-length transcript, as a function of reaction temperature. The major energy components of the Mfold ΔG are hydrogen bond energy and base pair stacking energy. These can be assumed to have a roughly linear relationship with transcript length. In order to make energies from different-length transcripts comparable, energies were normalized by computing a per-residue folding ΔG for each transcript and then multiplying that value by the global mean target length, for all transcripts considered from all organisms, of 851 bp. Average ΔG of secondary structure formation decreases with increasing temperature, but even at 65°C, the average ΔG of secondary structure formation for a full-length transcript is -98.2 kcal/mol (-27.9 kcal/mol when modeled as cDNA), meaning that the transcript is quite stable in that structure and a considerable energy input will be required to displace or melt the remaining structure. The trend in ΔG of secondary structure formation from the high-GC genome of *B. suis* to the low-GC genome of *L. lactis* is a decrease in overall stability. The average normalized ΔG of secondary structure formation for transcripts selected from the GC-balanced genome (*E. coli*) is near 70% of the average for *Brucella*, while the average ΔG for transcripts from the GC-poor genome (*L. lactis*) are even lower (30% at 52°C). However, even in the most GC-poor genome, stable target secondary structure in the single-stranded target is widespread.

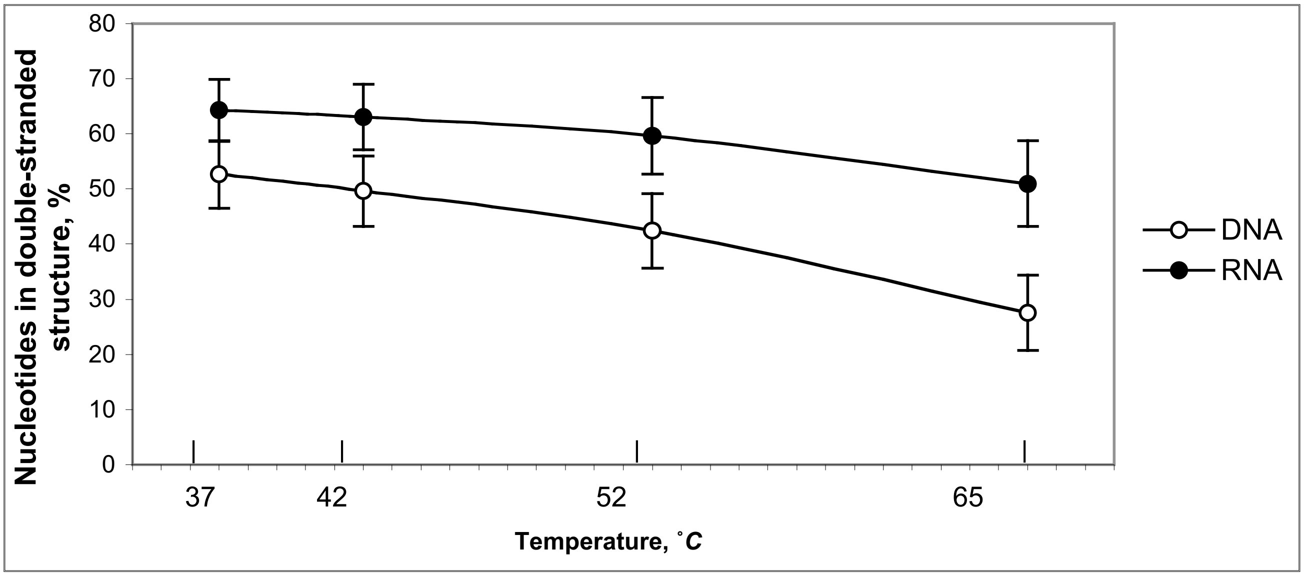


Figure 3
Fractional accessibility of nucleotides in the target. Fraction of the complete transcript classified as inaccessible due to the presence of stable structure in >50% of predicted conformations. Data shown are for 37, 42, 52 and 65°C simulations in *Brucella suis*.

Our results demonstrate that a significant fraction of nucleotide sites in the average target mixture, whether single stranded DNA or RNA, will be found in stable secondary structure under the hybridization conditions used in oligonucleotide microarray experiments, and will be relatively inaccessible for intermolecular interactions. Figure 3 shows the percentage of nucleotides that are in a double-helical state in at least 50% of the secondary structure conformations predicted by Mfold, at various reaction temperatures. The measure of accessibility used is the fraction of structures in which a nucleotide is found in a single-stranded conformation, when all optimal and suboptimal structures predicted are considered.

Extent and stability of target secondary structure

Figure 4 is a plot of the average ΔG of structure formation when shearing of the target molecule is simulated by dividing the target into overlapping 200, 100, and 50mer fragments. Shearing the target into smaller fragments destabilizes secondary structure, especially at very short fragment lengths. However, shearing does not eliminate occlusion of nucleotides by secondary structure, even in the shortest fragments examined. When a DNA target is modeled at 52°C, for example, the double stranded fraction decreases by only about 30% – from 41% to 29% – when the target is simulated as sheared into 50mer fragments. However, in hybridization experiments involving low copy number targets and longer oligos, creating extremely short target fragments may reduce or eliminate the signal on the chip, because the target can not be sheared specifically to present an unbroken hybridization site for the probe, and so some fragments will be created that match the probe only partially.

Interference of secondary structure with the hybridization site

Figure 5 shows the average percentage of nucleotides within a probe binding region in the target that are inaccessible, when different fractional accessibility cutoffs are used to classify the sites. Even when a relatively demanding criterion – double-strandedness in over 75% of optimal and suboptimal structures – is used to classify a nucleotide as inaccessible, an average of 21 ± 13% of nucleotides in the probe binding region are found in stable secondary structures at 65°C. Figure 6 shows a representative transcript and the challenge it presents to hybridization when modeled as full-length cDNA and fragments of various lengths.

Discussion

Lack of bioinformatics tools that incorporate experimentally validated biophysical properties of nucleic acids as a

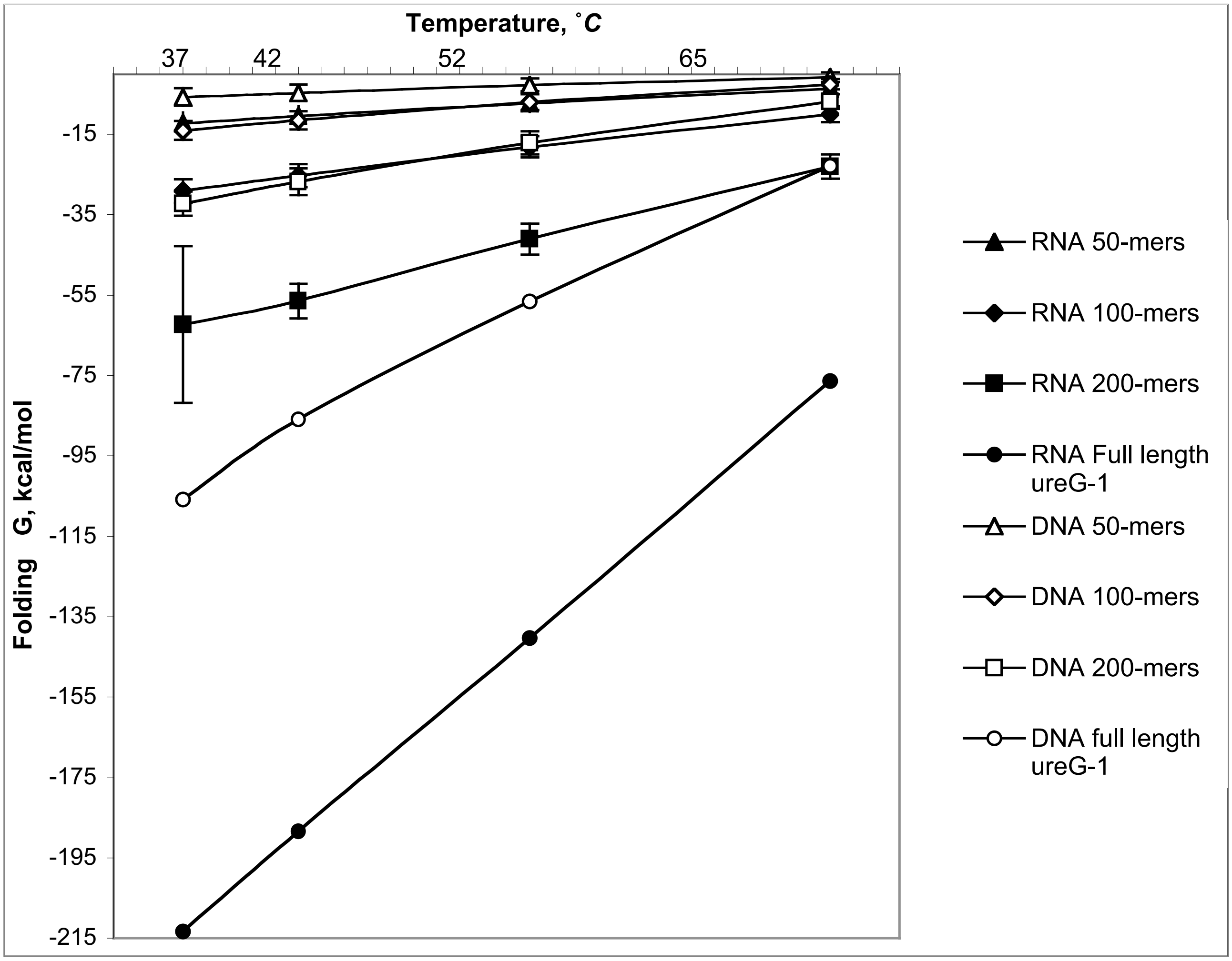


Figure 4
Stability of secondary structure in sheared fragments. Free energy change on secondary structure formation for the ureG-1 RNA transcript from *Brucella suis*. The transcript is modeled as sheared into fragments of length 200 nt, 100 nt or 50 nt; fragments are chosen starting at every 10th residue.

criterion for synthetic oligomer probe design is a major challenge for do-it-yourself microarray designers. One biophysical characteristic, which we predict will reduce the binding efficiency of microarray probes to their targets, is the propensity of long single-stranded DNA or RNA molecules to form stable secondary structure. 3-D structures such as hairpins and stacked regions have the potential to pre-empt target nucleotides, thus blocking regions of the target molecules from hybridizing to their intended probes. Prediction and thermodynamic analysis of secondary structure at a range of temperatures in full length target sequences, as well as in subsequences formed by *in silico* shearing, revealed the likely presence of stable secondary structures in both full-length target and sheared target mixtures. These structures do not convert completely to random coil with either increasing hybridization temperature, more extensive shearing, or both. These secondary structures may therefore compete with the intended target for effective probe annealing in a microarray experiment, resulting in a misinterpretation of the amount of target present in the sample.

Applying target secondary structure as a criterion in array design

Based on the results of this *in silico* experiment, secondary structure prediction in the target is being used to develop

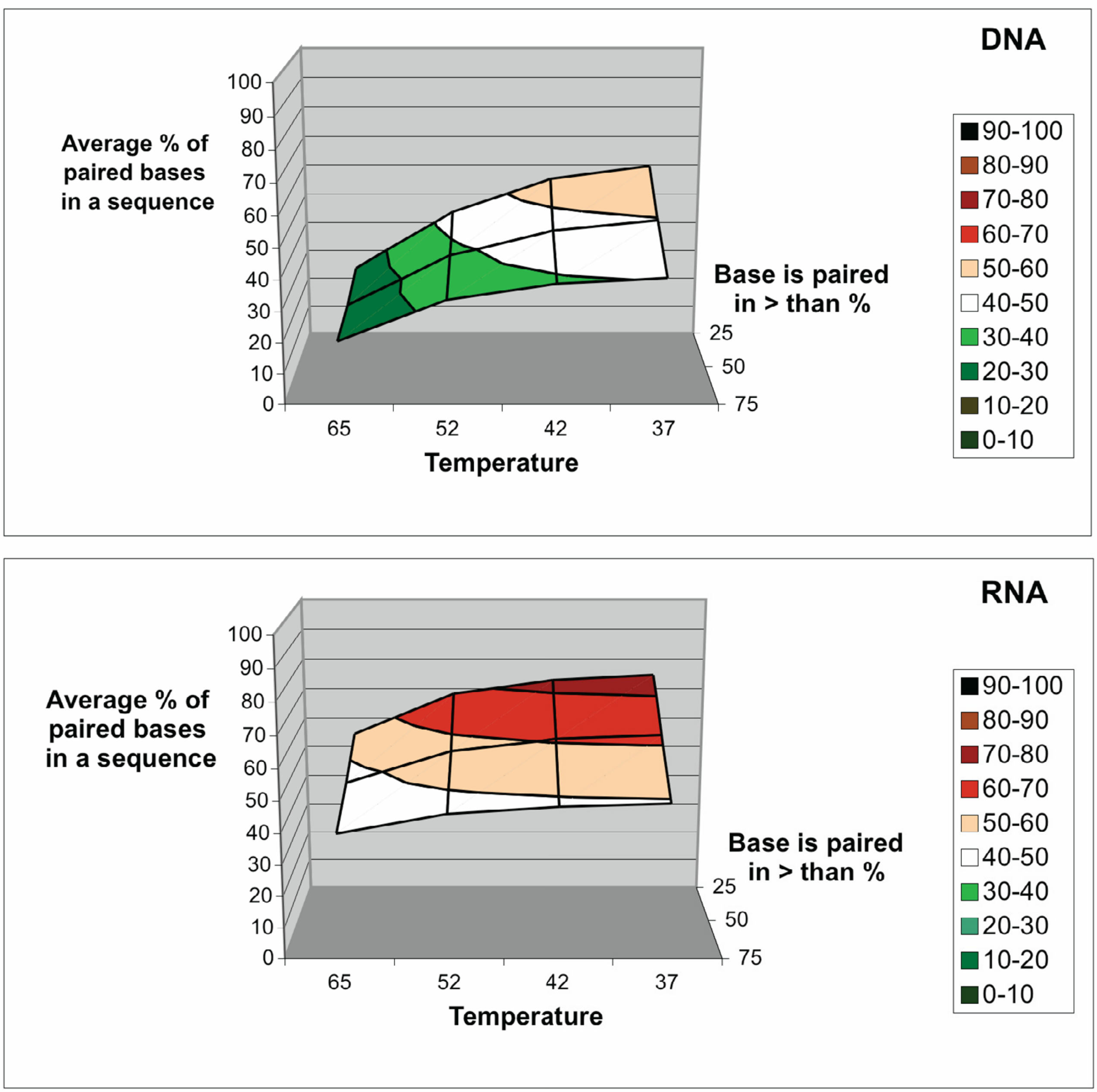


Figure 5

Accessibility of the probe binding site. Fraction of the average probe binding site in the *Brucella* genomic array that is found to be inaccessible at 37°, 42°, 52° and 65°C, for DNA or RNA target. Inaccessible sites are defined here using three different cutoffs for the fraction of structures in which the site is base-paired: 25%, 50%, and 75%.

a new criterion for oligonucleotide probe design. Our results from this modeling experiment demonstrate that the implicit assumption used until now – that eliminating probe secondary structure by avoiding self-complementarity eliminates target secondary structure as well – is valid only when the target and probe are of the same length. Use of target secondary structure as an explicit criterion will allow for masking or preferentially avoiding the

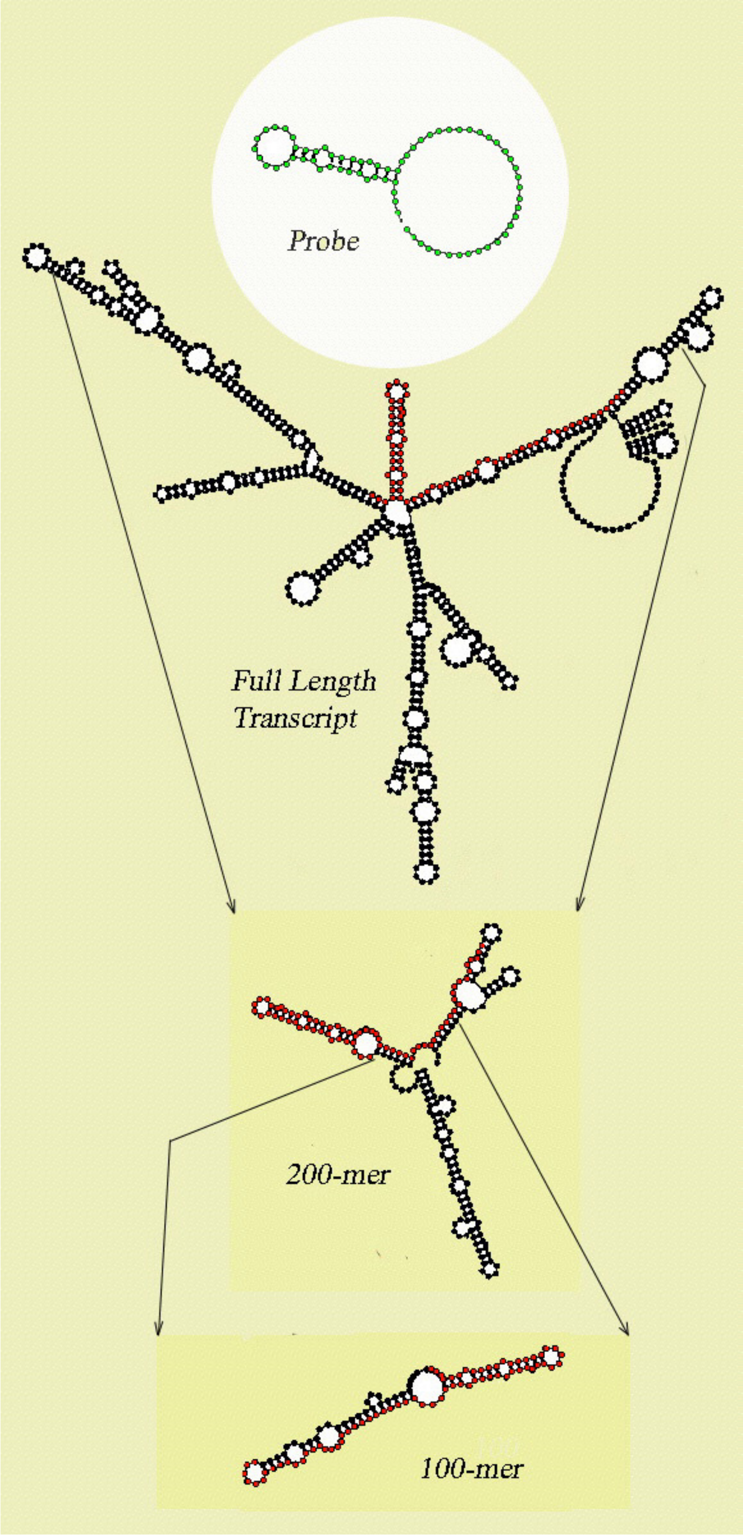


Figure 6

Structure in a binding site – full length target and sheared fragments. The position of a 70mer oligonucleotide probe (green) binding site (red dots) within a full-length optimal transcript structure, as well as examples of stable structure in 200mer and 100mer fragments which overlap the probe binding site. Corresponding ΔG values for these fragments modeled at 42° and 52°C are shown in Table 1.

regions of the target sequence in which base pairs are directly involved in secondary structure formation, to eliminate these regions from the sequence for the purpose of the search for the optimal probe.

In this study we have assigned accessibility scores to sites in the target sequence based only on the fraction of predicted structures within 5% of the energy optimum, in which a residue is found in a single-stranded conformation. While this measure is not too computationally intensive to compute, and can be applied to genome-scale problems using readily available software (Mfold), it is not the most physically rigorous definition of accessibility. By equally weighting each possible structure in the ensemble of optimal and suboptimal structures that a molecule can form, it is possible that secondary structure at some positions in the molecule is overcounted; bonds which form only in rare conformations are considered equal to bonds which are present in the lowest-energy structure. The program Sfold [16-18] assigns accessibility based on an ensemble-weighted average of secondary structure. The program RNAfold[19], part of the Vienna RNA package, implements McCaskill's partition function approach[20] to arrive at pairing probabilities for each pair of bases in the sequence, from which a summary per-base accessibility can be derived. These methods are more rigorous than MFold and we expected they might produce somewhat different results, although it has also been shown that predicted binding states from MFold optimal structures perform almost as well as SFold and RNAFold predictions when applied to molecules of known 3D structure [16].

When we compared MFold-based accessibility predictions for an individual transcript to those generated by SFold and RNAFold, we found that the difference in average predicted accessibility over an entire transcript is small. We computed accessibility for the transcript of human 1CAM-1, which has been mapped experimentally to determine its accessibility [21]. The average fractional accessibility derived from MFold results is about 3–4% greater than that predicted by RNAFold or SFold. Therefore use of this fractional accessibility measure will not impose an unnecessary constraint on the design process relative to other predictive approaches. The accessibility profiles calculated for ICAM-1 using each method are shown in Fig. 7. In each section of the figure, antipeak locations (having lower pairing probability and therefore likely to be more accessible) can be compared to the extendable sites detected by Allawi et al [21], which are indicated by green dots at the bottom of the plot. In each prediction, there are a number of apparently correct predictions and obvious errors, and it is not clear which method is yielding the best results at the residue level. A systematic, competitive test of these predictions against solution accessibility data gathered on various experimental platforms is called for, although available data sets for validation are still rare. In the absence of such validation, the MFold accessibility predictions are sufficient to predict the scope of the secondary structure problem in a genome-based array design, even if some details of the prediction are not correct. An experimental approach will eventually be required to determine which approach best represents the conditions of the microarray experiment.

Loop length and other considerations

In this study, we focused specifically on the DNA/RNA base pairs that are actively involved in hydrogen bond formation. We realize that other accessibility considerations will have to be added to the scoring scheme in practice. The structure of a long single stranded DNA or RNA molecule can contain many nucleotides that, while not part of a double-helical stem, remain inaccessible to hybridization due to their location inside small loops within the target secondary structure. A loop is a somewhat constrained structure as well, and the length at which it presents accessible sequence that favors hybridization has been shown to be on the order of 10 nucleotides and longer [22], while nucleotides found in shorter loops may be classifiable as inaccessible. However, there is a need for quantitative hybridization experiments that would elucidate how loops and loop-like structures in tethered long-oligo probe and target molecules affect the performance of assays, and we have chosen not to formulate a system for scoring the accessibility of single-stranded loop structures or weighting this criterion relative to the double-strandedness criterion until we have carried out some of these experiments.

Development of a target secondary structure criterion for oligonucleotide array design is expected to impose restrictions on the probe selection beyond the sequence similarity and melting temperature criteria that are currently used, especially in cases where short probe length restricts the annealing temperature used in the hybridization protocol to 22–37°. In the *B. suis* example, use of a low annealing temperature, e.g. 42°C which is the temperature used in some published 70-mer array experiments [9], would result in only about 30% of the average transcript being accessible for intermolecular hybridization, not counting 'free' bases found in short loops in secondary structures. There will be greater design latitude for experiments carried out at higher hybridization temperatures. Recommended hybridization temperatures for long synthetic oligomer arrays may prove to be closer to 65°C, when only 50% of a typical RNA transcript or 30% of the corresponding cDNA molecule remain inaccessible.

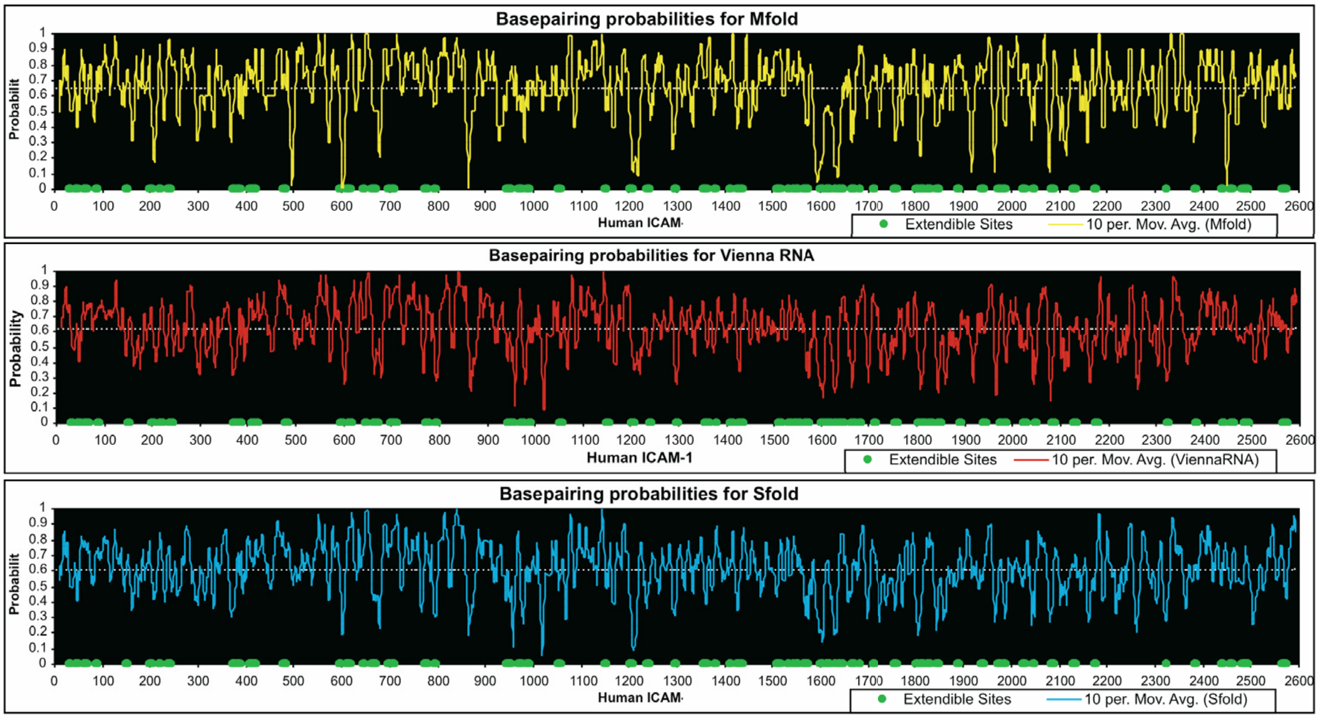


Figure 7

Accessibility prediction using three common methods. Pairing probabilities computed using RNAFold (top), MFold (middle) and SFold (bottom) for the human ICAM-1 transcript. Extendable sites detected by Allawi et al [21]

To shear or not to shear

We have shown here that while shearing reduces overall ΔG of secondary structure formation for individual molecules in the target solution, shearing does not in itself eliminate formation of secondary structure in single-stranded DNA or RNA. The question of whether shearing should be used for long oligomer arrays is still an open one. While some signal may be gained by reducing the stability of secondary structure in the target molecule, random shearing by its nature creates a mixture of targets that may have substantially different affinities. For instance, in a 300 nt transcript that is targeted by a 70mer oligonucleotide, there is nearly a one in four chance that a random break in the sequence will occur within the target site for which the probe is designed. Short fragments may present a substantially different binding site, and therefore have a different binding affinity, than the full-length transcript that is considered when the probe is designed. This is illustrated in Figure 8d, where binding of a 50mer sheared fragment to a 70mer probe leaves a dangling end in the probe. A break very close to one end or the other of the target site may create a target that still binds to the probe, though with reduced affinity; a break closer to the middle of the target site may produce fragments that bind partially to the probe, competing for binding with perfect matches.

The utility of experimentally validated biophysical criteria

In other experimental contexts where hybridization is critical to success, the impact of secondary structure in single stranded polynucleotides on results has been recognized and is now being systematically studied (18–21). Intramolecular folding of mRNAs is so extensive that only 5–10% of most transcripts is accessible to binding of complementary nucleic acids; however the modeling of long molecules has not proven to give very accurate binding predictions [23-25]. In fact, array-based screens have been utilized to empirically select oligonucleotides that bind effectively to transcripts for siRNA experiments [23,26]. Several studies have demonstrated that, at 37°C and 0 mM Mg2+ oligonucleotides of length >20 yield good binding/RNAseH digestion at low concentrations relative to shorter oligonucleotides (30 nM vs 300 nM compared) and found that microarray binding was a good predictor of siRNA activity despite the 3' tethering and 1M NaCl used in array experiments vs siRNA experiments [26]. Systematic "scanning" of mRNA sequences with libraries of short oligos [27] has also been shown to be successful in locating sites for siRNA targeting; however, such methods are likely to become extremely expensive if applied to the large number of targets in a microarray design. We have begun to develop an experimental approach to this problem, in which structure predictions like those used in this study are experimentally evaluated to determine whether the structures we can predict using existing modeling approaches will detectably affect signal in the microarray context.

Conclusion

The results of the current study suggest a significant role for target secondary structure in hybridization to oligonucleotide arrays, which will warrant further investigation. Oligonucleotide probe binding sites in a significant fraction of transcripts are found in double-stranded conformations even in cases where self-complementarity was avoided during the probe design process. We find that at 52°C, for example, approximately 57% of probes designed for *Brucella* had binding sites in the target which were predicted to contain a stretch of unpaired bases of at least 14 nt in length; at 65°C, that fraction increased to 93%. Based on these findings we would expect that at 52°C only 57% of our probes would encounter optimal conditions for hybridization and therefore would demonstrate the expected behavior in the experiment, where intensity is expected to scale with target concentration. We predict that the remaining probes, which have shorter, or no, accessible sequences, will exhibit modified binding behavior, and we plan to conduct experiments to characterize this behavior. We have shown conclusively that avoiding self-complementarity in the probe when designing an oligonucleotide array is insufficient to eliminate secondary structure from the binding site in the target. By combining the procedure for systematic computational assessment of transcript accessibility described in this study with selective experimental validation of the impact of predicted accessibility on hybridization, we will develop a useful criterion for avoiding troublesome secondary structure when designing microarray targets.

Methods

Prediction and thermodynamic analysis of secondary structure was performed for all protein-coding gene transcripts predicted from 3264 CDSs in the *Brucella suis* 1330 genome. *Brucella suis* has a relatively high (57%) genomic GC content. *Brucella suis* was chosen for this experiment because our collaborators have previously acquired a custom synthetic oligomer microarray for this organism, developed using standard oligo array design software, and we have access to both target sequences and to a set of unique probe sequences that define the interaction sites for which expression results have been obtained by the laboratory.

In order to determine whether Brucella sequences form atypical structures we randomly picked and analyzed 50 gene coding sequences from a compositionally balanced genome (*Escherichia coli*), and 50 from the GC-poor genome of the nonpathogenic AT-rich gram-positive bacterium *Lactococcus lactis* (35% genomic GC content). The *Brucella suis* genes ranged in length from 90 to 4,803 bp, with an average transcript length of 851 bp. The *E. coli* genes ranged in length from 140 to 2,660 bp, with an average transcript length of 792 bp. The range of GC content in the genes chosen was 37% to 57% with an average value of 50%, which is reasonably representative of the *E. coli* genome. The *L. lactis* genes chosen ranged in length from 140 to 2,730 bp, with an average transcript length of 765 bp., and ranged in GC content range from 30% to 42% with an average value of 35%.

Microarray design

70-mer probes for each *Brucella suis* target were previously designed (Stephen Boyle, personal communication) using ArrayOligoSelector (pick70) [10]. ArrayOligoSelector uses sequence uniqueness, self-complementarity, and sequence complexity as criteria but does not explicitly evaluate ΔG of secondary structure formation for the probe. 72% of the probes designed using this method were found to contain secondary structures with melting temperatures greater than 65°C, and 10% contained secondary structures with melting temperatures greater than 80°C. The Brucella probes defined the interaction sites within the target transcripts for which structural accessibility was evaluated.

Secondary structure prediction

Probe and transcript secondary structure were predicted using the Mfold 3.1 software package [28,29]. Mfold identifies the optimal folding of a nucleic acid sequence by energy minimization and can identify suboptimal foldings within a specified energy increment of the optimum as an approach to modeling the ensemble of possible structures that a single-stranded nucleotide molecule can assume. We modeled secondary structure in the single-stranded target, modeling the target both as DNA and as RNA, at a range of temperatures which is inclusive of hybridization temperatures commonly used in microarray protocols: 37°C, 42°C, 52°C and 65°C. The modeling conditions were chosen within the allowed settings of Mfold to approximate a microarray experiment: solution conditions of 1.0 M sodium concentration and no magnesium ion were used. The free energy increment for computing suboptimal foldings, ΔΔG, was set to 5% of the computed minimum free energy. The default values of the window parameters, which control the number of structures automatically computed by Mfold 3.1, were chosen based on the sequence length. Free energy changes on

Table 1: Stability of a sample transcript – full length target and sheared fragments Folding ΔG of target transcript and fragment molecules shown in Figure 8, at hybridization temperatures commonly used for long oligomer arrays.

Molecule	**^{2}G, kcal/mole**			
	42°C		**52°C**	
	DNA	**RNA**	**DNA**	**RNA**
70-mer Probe	- 6.8	N/A	- 4.2	N/A
Full Length Target	- 85.9	- 188.4	-56.6	- 140.2
200-mer sheared Target	- 25.5	- 58.6	-15.9	- 41.6
100-mer sheared Target	-14.2	- 25.7	-9.6	-18.0
50-mer sheared Target (not shown)	- 6.1	-10.5	- 4.2	-7.3

formation of secondary structure were extracted from the Mfold output.

Accessibility calculation

Accessibility in folded single-stranded DNA or RNA has recently begun to be addressed in a few experimental studies, mainly with the goal of targeting appropriate sites for RNAi. Because the structure of single-stranded nucleotide molecules is much more dynamic than that of proteins, with each molecule likely to exist in an ensemble of structures, and because the 3D structure of these molecules is rarely known, there is not yet a consensus representational standard of per-residue accessibility for single-stranded nucleic acids. Ding et al. [17,18] implement probability of single-strandedness, when the weighted ensemble of likely structures is taken into account, as an accessibility criterion. However, use of their Sfold server, with batch jobs limited to 3500 bases, is not currently practical for a genome-scale survey of accessibility. Another approach to accessibility prediction is McCaskill's partition function approach [20] which can be used to compute base pair probabilities and summary pairing probability for any base. This approach is implemented in RNAFold [19], a component of the Vienna RNA package.

In this study, we chose to use the less physically rigorous approximation of probability of single strandedness as a simple fraction of predicted optimal and suboptimal structures in which a residue is found to be part of a single stranded structure, as computed by Mfold. Accessibility scores derived from MFold predictions have been used in limited studies of RNA structure focused on hammerhead ribozymes[30], antisense and siRNA targeting [22,31] and have been shown to be predictive in cases where some experimental measure of accessibility has been made[32]. While MFold-derived accessibility scores may not be completely optimal, they have been used with reasonable success to predict accessibility in the siRNA targeting context, and so we use MFold here.

Shearing simulation

Random shearing of the target mixture is an approach that is often offered as a solution for the problem of target secondary structure. The actual content of a sheared mixture of DNA or RNA fragments is complex. Shearing breaks the molecule not in predictable locations, but in random locations that give rise to a distribution of fragments around an average fragment length. In order to simulate the effects of different degrees of shearing on structure formation and stability in a transcript, we picked fragments of 200, 100, or 50 bases in length, choosing the start position via a sliding window of 10 bases. Secondary structure prediction for all fragments derived from every transcript in the B. suis genome is computationally intensive and produces an extremely large amount of output. Since our initial goal was to determine how much the method would affect the number and type of secondary structures probes would be expected to bind the shearing simulation was performed for fragments derived from the 300 bp Ure-1A gene of *B. suis*. Secondary structure and thermodynamics were computed for each of these fragments individually.

Authors' contributions

VGR participated in the design of the study, carried out the simulations and analysis, and drafted the manuscript. JWW participated in the design of the study and helped to draft the manuscript. CJG conceived of the study, participated in its design, coordinated the research and analysis, and drafted the manuscript.

Acknowledgements

The authors would like to acknowledge the Fralin Biotechnology Center for research startup support (CJG) and the Department of Biology for graduate assistantship support (VGR).

References

1. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL: **Expres-**

sion monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996, **14(13):**1675-1680.

2. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270(5235):**467-470.
3. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ: **Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays.** *Nucleic Acids Res* 2000, **28(22):**4552-4557.
4. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.** *Nat Biotechnol* 2001, **19(4):**342-347.
5. Ramakrishnan R, Dorris D, Lublinsky A, Nguyen A, Domanus M, Prokhorova A, Gieser L, Touma E, Lockner R, Tata M, Zhu X, Patterson M, Shippy R, Sendera TJ, Mazumder A: **An assessment of Motorola CodeLink microarray performance for gene expression profiling applications.** *Nucleic Acids Res* 2002, **30(7):**e30.
6. Chou CC, Chen CH, Lee TT, Peck K: **Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression.** *Nucleic Acids Res* 2004, **32(12):**e99.
7. Shi SJ, Scheffer A, Bjeldanes E, Reynolds MA, Arnold LJ: **DNA exhibits multi-stranded binding recognition on glass microarrays.** *Nucleic Acids Res* 2001, **29(20):**4251-4256.
8. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, Johnston R: **An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression.** *Nucleic Acids Res* 2001, **29(8):**E41-1.
9. Wang HY, Malek RL, Kwitek AE, Greene AS, Luu TV, Behbahani B, Frank B, Quackenbush J, Lee NH: **Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays.** *Genome Biol* 2003, **4(1):**R5.
10. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL: **Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray.** *Genome Biol* 2003, **4(2):**R9.
11. Chou HH, Hsia AP, Mooney DL, Schnable PS: **PICKY: oligo microarray design for large genomes.** *Bioinformatics* 2004.
12. Nielsen HB, Wernersson R, Knudsen S: **Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays.** *Nucleic Acids Res* 2003, **31(13):**3491-3496.
13. Rouillard JM, Zuker M, Gulari E: **OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach.** *Nucleic Acids Res* 2003, **31(12):**3057-3062.
14. Tolstrup N, Nielsen PS, Kolberg JG, Frankel AM, Vissing H, Kauppinen S: **OligoDesign: optimal design of LNA (locked nucleic acid) oligonucleotide capture probes for gene expression profiling.** *Nucl Acids Res* 2003, **31(13):**3758-3762.
15. Nguyen HK, Southern EM: **Minimising the secondary structure of DNA targets by incorporation of a modified deoxynucleoside: implications for nucleic acid analysis by hybridization.** *Nucl Acids Res* 2000, **28(20):**3904-3909.
16. Ding Y, Chan CY, Lawrence CE: **Sfold web server for statistical folding and rational design of nucleic acids.** *Nucleic Acids Res* 2004, **32(Web Server issue):**W135-41.
17. Ding Y, Lawrence CE: **Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond.** *Nucleic Acids Res* 2001, **29(5):**1034-1046.
18. Ding Y, Lawrence CE: **A statistical sampling algorithm for RNA secondary structure prediction.** *Nucleic Acids Res* 2003, **31(24):**7280-7301.
19. Hofacker ILFWSPFBSTMSP: **Fast folding and comparison of RNA secondary structures.** *Monatshefte f Chemie* 1994.
20. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29(6-7):**1105-1119.
21. Allawi HT, Dong F, Ip HS, Neri BP, Lyamichev VI: **Mapping of RNA accessible sites by extension of random oligonucleotide libraries with reverse transcriptase.** *Rna* 2001, **7(2):**314-327.
22. Scherr M, Rossi JJ, Sczakiel G, Patzel V: **RNA accessibility prediction: a theoretical approach is consistent with experimental studies in cell extracts.** *Nucleic Acids Res* 2000, **28(13):**2455-2461.
23. Sohail M, Akhtar S, Southern EM: **The folding of large RNAs studied by hybridization to arrays of complementary oligonucleotides.** *Rna* 1999, **5(5):**646-655.
24. Lima WF, Monia BP, Ecker DJ, Freier SM: **Implication of RNA structure on antisense oligonucleotide hybridization kinetics.** *Biochemistry* 1992, **31(48):**12055-12061.
25. Michel F, Westhof E: **Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis.** *J Mol Biol* 1990, **216(3):**585-610.
26. Bohula EA, Salisbury AJ, Sohail M, Playford MP, Riedemann J, Southern EM, Macaulay VM: **The efficacy of small interfering RNAs targeted to the type 1 insulin-like growth factor receptor (IGF1R) is influenced by secondary structure in the IGF1R transcript.** *J Biol Chem* 2003, **278(18):**15991-15997.
27. Zhang HY, Mao J, Zhou D, Xu Y, Thonberg H, Liang Z, Wahlestedt C: **mRNA accessible site tagging (MAST): a novel high throughput method for selecting effective antisense oligonucleotides.** *Nucleic Acids Res* 2003, **31(14):**e72.
28. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31(13):**3406-3415.
29. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.** *J Mol Biol* 1999, **288(5):**911-940.
30. Amarzguioui M, Brede G, Babaie E, Grotli M, Sproat B, Prydz H: **Secondary structure prediction and in vitro accessibility of mRNA as tools in the selection of target sites for ribozymes.** *Nucleic Acids Res* 2000, **28(21):**4113-4124.
31. Amarzguioui M, Prydz H: **An algorithm for selection of functional siRNA sequences.** *Biochem Biophys Res Commun* 2004, **316(4):**1050-1058.
32. Kretschmer-Kazemi Far R, Sczakiel G: **The activity of siRNA in mammalian cells is related to structural target accessibility: a comparison with antisense oligonucleotides.** *Nucleic Acids Res* 2003, **31(15):**4417-4424.

VITA

Vladyslava Ratushna was born on April 26, 1976 in Kyiv, Ukraine. In 1993 she finished the local high school with the Golden Medal Award, and was admitted to the National University of “Kiev-Mohyla Academy”(NaUKMA). As a sophomore, she participated in a one year exchange program with the University of Texas, Austin (UT Austin). Her undergraduate research interests concerned the effect of the radionuclide incorporation on a structure of rat bone tissue. In 1998 she graduated from NaUKMA with a bachelor degree in Natural Sciences with a major specialization in Biology and a minor in Ecology. After that she entered a graduate program at the Department of Biophysics, Biochemistry and Molecular Biology at Iowa State University (ISU), where she worked in a *Zea mays* molecular genetics lab. Vladyslava transferred from Iowa to Virginia Tech and conducted bioinformatics research in Dr.Gibas’s lab at the Department of Biological Sciences. She will continue her study of the effect of the target secondary structure on the quality of the microarray experiments at the University of North Carolina, Charlotte.