Flanking sequence effects on oligonucleotide hybridization.

D. Andrew Carr[1], Saeed Koshnevis[2], Donald Kolva[1], and Jennifer W. Weller[2]

(1)Accelerated Technology Laboratories, Inc 496 Holly Grove School Road West End, NC 277376 USA.
(2) Bioinformatics/Computer Science, University North Carolina Charlotte,
9201 University City Blvd, Charlotte, NC 28223 USA.

Accelerated Technology LABORATORIES INC.

SeqNFind

UNC CHARLOTTE

UNC CHARLOTTE COLLEGE OF COMPUTING AND INFORMATICS BIOINFORMATICS research center

Abstract

Microarray platforms measure both sequence content and prevalence of sequence by leveraging target hybridization to immobilized, high-concentration oligonucleotides. Probe design algorithms optimize for unique, short regions of high sequence similarity, usually 24-70nt in length. Gain of signal from untargeted genomic sequences ('cross-hybridization') is minimized through a variety of pattern-matching algorithms specific to the probe length, with detectable signal predicted to arise from a small number of mismatches but not loops. Loss of signal from competing structures is predicted using folding algorithms. It is generally assumed that the target structure affects heteroduplex binding only if the probe binding site is obstructed.

In this study, we examine the effects of the full length of a target's sequence on microarray hybridization under standard assay conditions. Two factors were tested: whether a stable duplex results when loops are present in the target in the binding site, and whether target structure in flanking sequences, beyond the probe binding site, affects the stability of the heteroduplex. Selecting a small subset of 33mer probes from the Affymetrix SNP6.0 Array we generated a pool of gapped, but predicted to be stable, sequence complements in the human genome (reference version 36.3) using the SeqNFind™ platform. Centered on the heteroduplex regions, sets of increasing target length were generated, to mimic a DNAse I digestion. The products were modeled using OMP™ to calculate the free energies of the major species, and the predicted percent bound target. To demonstrate that the modeled properties are experimentally important (affect measurements) several targets and probes were tested on microarrays.

Introduction

What is cross-hybridization?

Cross-hybridization is the annealing of an unintended target to a probe – it is specific but does not report on the expected locus. Partially matched targets will also bind probes, albeit with lower energy – however if the concentration is high significant signal can still result. Hybridization conditions and concentration of species dictate whether partial matches become detectable. The stable binding of an unintended target leads to interpretation errors.[1]

Relevance to SNP 6.0 arrays?

With a smaller dynamic range than expression arrays, SNP studies rely on a one to one relationship between the genome sequence at a locus and its probe and very subtle sequence differences. Influence from an unintended location can result in the false identification of an allele. Our goal is to understand how wet-lab methods like PCR and DNAse I, computational methods, and individual genome structural differences affect the accuracy, specificity and sensitivity of allele calls.

Gibbs Free Energy (ΔG) is one measure of structural stability. Smaller ΔG (more negative) is more tightly bound, and therefore a more stable structure. Work by [2] showed that for one version of Affymetrix 25-mer arrays the ΔG was [-42.08 to -18.99kcal/mol] with a mean of -29.15kcal/mol. The modeling was to simple heteroduplexes across the intended complementary region only.

Sequence algorithms and structures

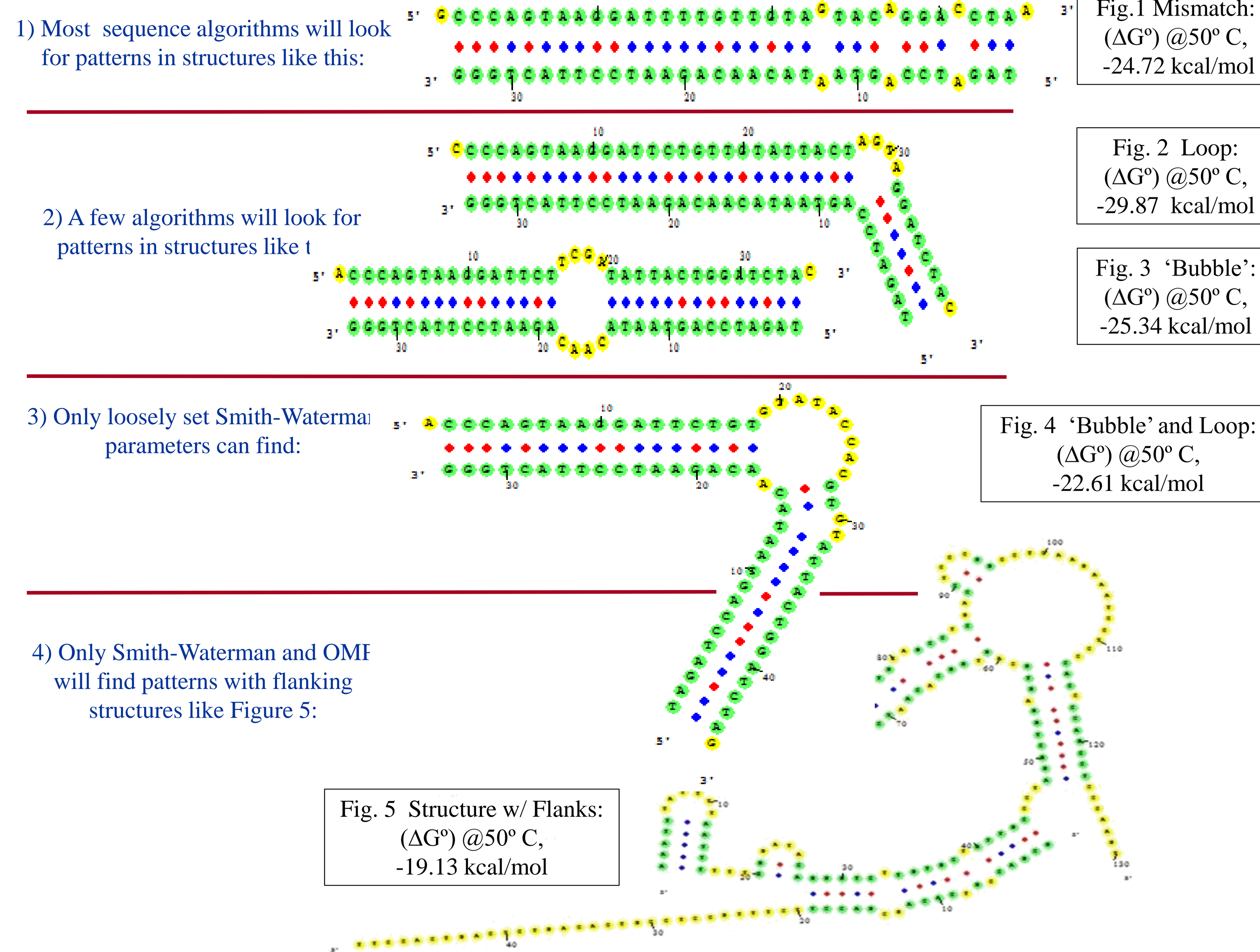


Fig.1 Mismatch: (ΔG^o) @50° C, -24.72 kcal/mol

Fig. 2 Loop: (ΔG^o) @50° C, -29.87 kcal/mol

Fig. 3 'Bubble': (ΔG^o) @50° C, -25.34 kcal/mol

Fig. 4 'Bubble' and Loop: (ΔG^o) @50° C, -22.61 kcal/mol

Fig. 5 Structure w/ Flanks: (ΔG^o) @50° C, -19.13 kcal/mol

Do the target sequence flanks surrounding a "probed" region effect the signal intensity produced by a micro-array?

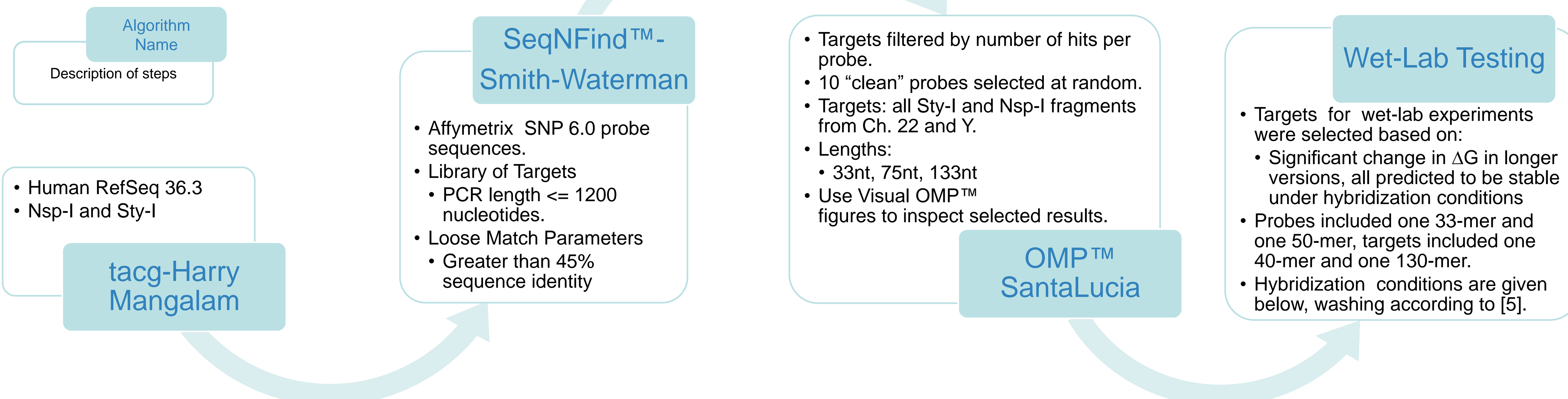


Fig 6. Target Sequences:

T_643_40_CTAGGCTGGAGTGCAGTGGCACAATCATGGAGTGCAGCCT

T_643_130 AAATTTATTTTAATTTTTTTGGATACAGGGTCTTGTGCTGTTGCCCTAGGCTGGAGTGCAGTGGCACAATCATGGAGTGCAGCCTCGACCTCCCGGCCTCAAGAAATCCTCCCACCCCAGCCTCCCAAGT

Assay Design

- A 33mer probe P643 (SNP_A-8475541, rs2356626) was selected as an example based on the SeqNFind™ and OMP™ modeling.
- Print 6 slides with [Probe] = 25µM and duplicate arrays top/bottom.
- Hybridize one of two nested targets (L = 40nt or 130nt); sequences are from HuRef 36.3 (Figure 6).
- OMP™ simulations predict that the targets yield equal signal under the same hybridization conditions (Table 1) and that both targets bind with the same bases to the probe (Figure 7).
- OMP™simulations predict that the longer target has secondary structure flanking the heteroduplex site lacking in the duplex with the shorter target.

Array Conditions

Assay Temp	45ºC
Na^+	0.056M
DMSO	0.96%
TMACl	3.68M
pH	5.6

Array Layout

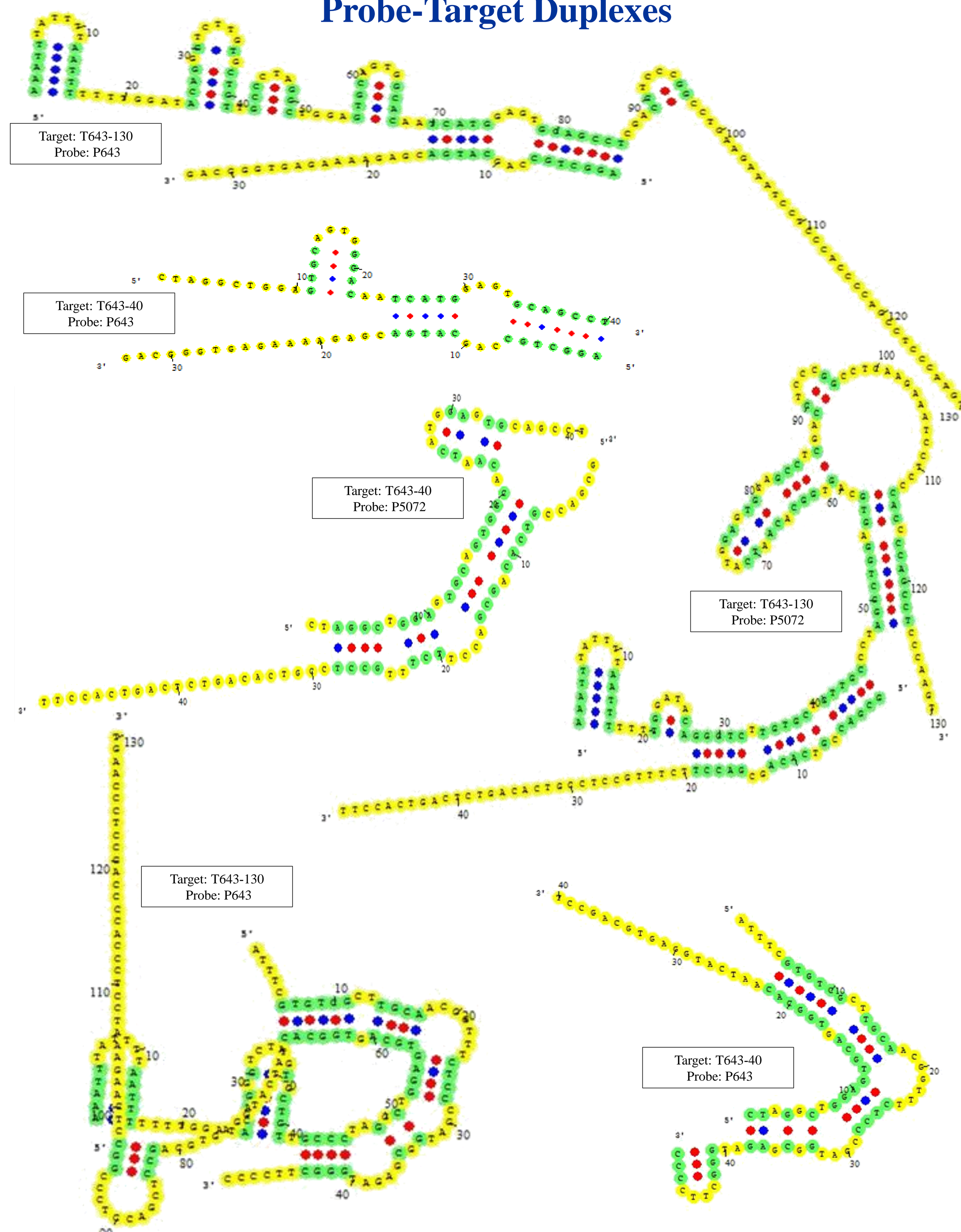
⇨Buffer
⇨Negative Control (GALM-Br0 857) 5uM
⇨Probe SNP 643-33 (25uM)
⇨Sentinel (5076 cy3) 5uM

⇨Buffer
⇨Negative Control (GALM-Br0 857) 5uM
⇨Probe SNP 643-33 (25uM)
⇨Sentinel (5076 cy3) 5uM

OMP Results

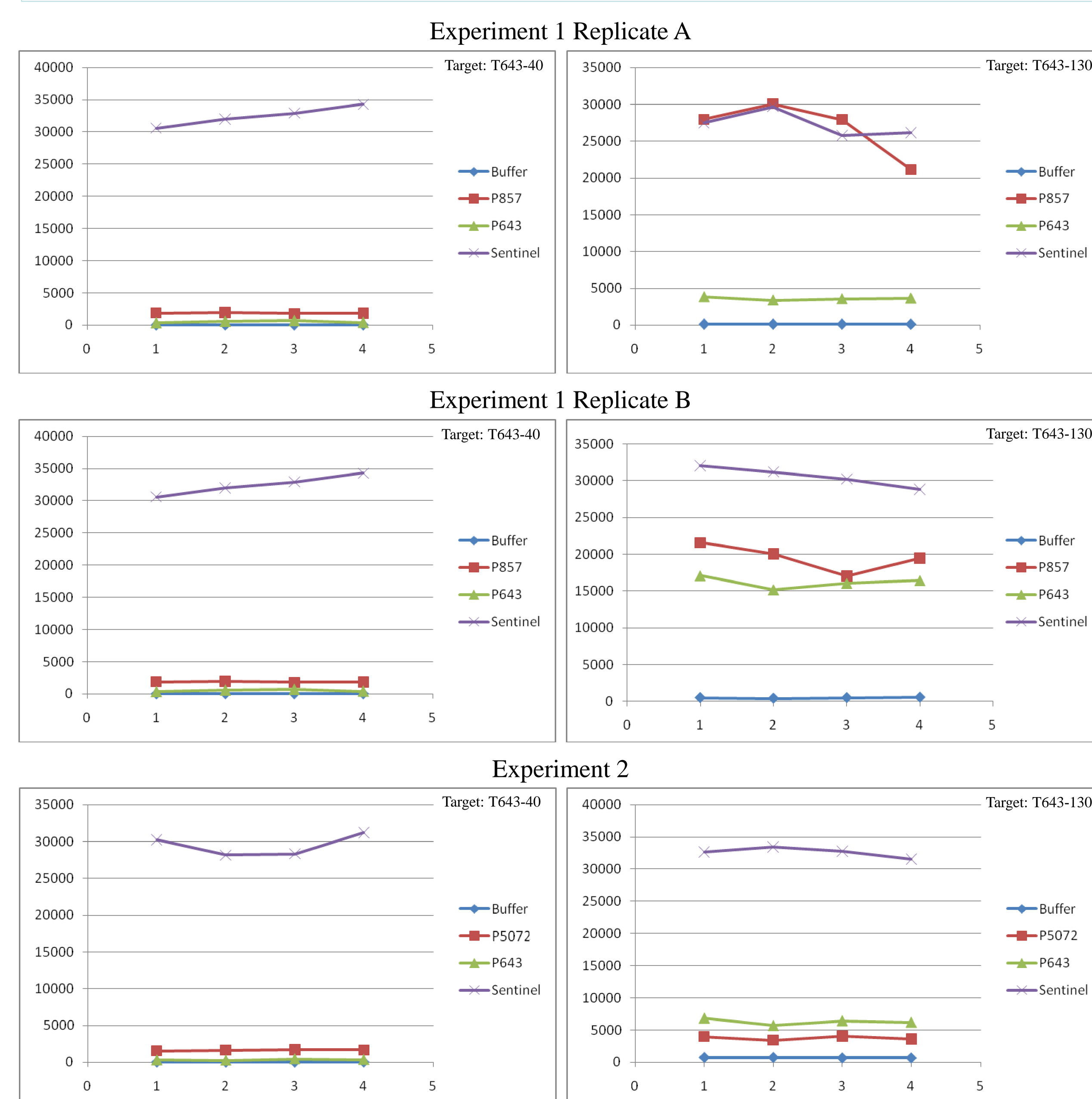
Target	Probe	ΔG (45°C)	Percent Bound
T_643_40	P643	-14.64	19.1
T_643_130	P643	-19.92	37.5
T_643_40	P5072	-7.17	0
T_643_130	P5072	-19.13	0
T_643_40	P857	-8.78	0
T_643_130	P857	-13.19	0

Probe-Target Duplexes



Microarray Results

- Experiment 1A and 1B:
  - P643 does has measurable signal for the 130mer but not for the 40mer.
  - P847 (which has a much lower predicted ΔG and no predicted percent bound) gives measurable signal with the 40mer target although this is close to the arrays minimum, the 130mer target length shows solid signal.
  - P847 although not predicted to show signal via modeling performs very well.
- Experiment 2 :
  - P5072 Based on modeling should not show signal, but does due to the flank stabilization.



Software

Visual OMP™ www.dnasoftware.com
SeqNFind™ www.atlab.com
tacg by Harry Mangalam

References

[1] M. Thompson and N. Woodbury; Thermodynamics of Specific and Nonspecific DNA Binding by Two DNA-Binding Domains Conjugated to Fluorescent Probes. *Biophysical Journal*, 81:1793-1804. 2001.
[2] K. Smith and M. Hallett; Towards Quality Control for DNA Microarrays. *Journal of Computational Biology,* 11(5): 945-970. 2004.
[3] SantaLucia J Jr.; Physical principles and visual-OMP software for optimal PCR design. *Methods Mol Biol*, 402:3-34. 2007.
[4] E. G. Shpaer; M. Robinson; D. Yee; J. D. Candlin; R. Mines & T. Hunkapiller; Sensitivity and selectivity in protein similarity searches: a comparison of Smith-Waterman in hardware to BLAST and FASTA. *Genomics,* 38:179-191. 1996.
[5] AE Pozhitkov, I Boube, MH Brouwer, PZ Noble "Beyond Affymetrix arrays: expanding the set of known hybridization isotherms and observing pre-wash signal intensities" *NAR* 38(5): e28. 2010.
[6] JC Wu, DP Gardner, S Ozer, RR Gutell, P Ren "Correlation of RNA Secondary Structure Statistics with Thermodynamic Stability and Applications to Folding" *J. Mol. Biol.* 391(4):769-783. 2009.
[7] H. Mangalam; tach –a grep for DNA. *BMC Bioinformatics*, 3:8 2002.

Conclusions

1. Neither free energy (ΔG) nor percent bound correlated well with the observed binding of targets to probes.
2. Under standard hybridization conditions it is possible to obtain a significant, detectable signal when the series of consecutive bases on the probe involved in the duplex is only 6 if you count non-canonical pairs and 8 in the target (where loops intervene).
3. Hairpin structure in the flanking regions of the target stabilize the duplex.
4. There is no software currently available that efficiently models these factors, and it may be necessary to update energy parameters also [6].