A Computational Study of Cross-hybridizing Stable Loop Structures in Oligonucleotide Sequences.

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

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

Abstract

Many genomics platforms that produce sequence and expression data use hybridization of a short nucleotide fragment to a longer target fragment as an essential step somewhere in the process. Uniqueness of the duplex is often a pre-requisite for success of the assay. The search space of a complete genome against a short complementary match is very large, so most platform design tools simplify the problem by limiting the oligo search space to contiguous regions and near perfect matches, and using heuristic, incomplete search algorithms.   Due to the flexible nature of the single nucleotide strand backbone, loops and bubbles can occur, causing stable hybridization between species with considerably less sequence homology then search tools commonly allow. In this study we've calculated the stability of such structures and then looked for examples in the human genome. We start with a hand crafted sequence/structure set of probe-target complements, and determine the free energy of the structures using Visual OMP^{TM}.   The subset of stable templates formed the basis of a search space, using as a sequence core a subset of the SNP probes (33-mers) on the Affymetrix SNP6.0 platform and Human reference genome build 36.3. The short probe lengths and gapped sequence makes this resemble a splice form detection problem, so BLAST-like algorithms are not suitable and a custom algorithm was developed. With adjustment of the algorithmic parameters in SeqNFind^{TM}, a pool of potential targets was constructed. Where examples were found free energies were again calculated. Conclusion: the potential for cross-hybridization is considerably greater than is currently recognized when loop structures, etc. are included.

Introduction

What is cross-hybridization?

Cross hybridization is the annealing of a partially matched complementary oligonucleotide probe to a "targeted" strand of DNA. In the case of Microarrays and PCR experiments the binding of an unintended sequence to a designed target can lead to errors interpreting signals or selecting amplification sites, respectively.[1]

Why is it important to SNP 6.0 arrays?

SNP studies rely on a one to one relationship between the genome sequence at a locus and its probe. Influence from an unintended location can result in the false recognition of a SNP (allele). Our long term goal is to examine how wet-lab methods, computational methods, and individual genome structural differences affect the accuracy, specificity and sensitivity of SNP calls.

Alignment patterns as predictors of cross-hybridization

Calculating the free energy of binding for all 1.8 million 33mer Affymetrix SNP 6.0 probes against ~7 billion locations in the reference genome is currently computationally intractable. This study uses alignment patterns produced by affine gap algorithms as filters to reduce the computational space and locate thermodynamically sound cross hybridizing locations within RefSeq build 36.3 of the human genome.

Software Tools Utilized

Visual OMP[TM]   www.dnasoftware.com
SeqNFind[TM]   www.atlab.com

Gibbs Free Energy

- ❖ Gibbs Free Energy (ΔG) is used as a measure of DNA strand to strand binding. Smaller ΔG (more negative) is more tightly bound, and therefore a more stable structure. Work by Smith and Hallett 2004, showed that for 25mer Affymetrix HuGeneFL™ chip C arrays ΔG for stable binding structures ranged between -42.083 and -18.993, with a mean of -29.1549 (Smith and Hallett Fig2. [2004]), with a maximal range of -45.083 to -15.603 (Smith and Hallett Fig1. [2004]). [2]

- ❖ Visual OMP[TM] This study uses Visual OMP[TM], a tool based on the nearest neighbor approach developed to calculate the (ΔG) on sequences. [3]

- ❖ (ΔG) Threshold: As a conservative starting point, a (ΔG) threshold of -22kcal/mol was selected to begin initial design work. Results shown provide counts for the entire range from -15kcal/mol and below. [2]

Pattern searching algorithms

- ❖ Algorithms: Not all Affine/Gap algorithms and implementations are equal. Some algorithms such as BLAST, BLASTN and those based on a heuristic approach are known to provide incomplete results sets. [4] Traditionally, heuristics were/are used to compensate for the complexity of the search space and the speed of the computer systems. Advances in computational hardware have allowed deployment of algorithms that yield more complete solutions in real time. In this study we used a new, optimized Smith-Waterman, 'SeqNFind[TM]', to search the genomic sequence. SeqNFind[TM] is based on a sliding window approach rather than heuristic seeding.

- ❖ Scoring parameters: Just as a shift in pattern matching algorithms can change the result set found, altering the scoring parameters used by an algorithm affects the results. In this study we were looking for balance between mismatch placement and indels (gaps) while staying within the energetic boundaries of what yields likely cross-hybridizing locations. Numerous iterations

were performed, to tune the system. A two-pass method, i.e. genomic scans were run twice with two different sets of parameters, was used to generate the data shown.

Step 1: Creation of an example structure set

❖ To begin the search for possible cross-hybridizing gapped 33mer structures, it was necessary to build a suite of examples. The initial structures were designed on the base sequence of a 33mer SNP probe (SNP A-8280034), randomly chosen from the Affymetrix SNP6.0 probe set. A series of sequential variants was constructed by hand, to represent structural categories incorporating types of 'gaps'. We generated hundreds of structures.

❖ Structure Prototypes: duplexes that contain alignment mismatches that result in small bubbles of approximately 1-4bp in length (Fig. 1, Fig. 3-6); Loop structures where the sequence to which the probe is hybridizing has insertions (Fig 2., Fig 6.); Structures that are open at least on one end.

❖ Threshold: a threshold of (ΔG°), for 50ºC, -22.00kcal/mol was used as a cutoff. Parameters for hybridization were extracted from the Affymetrix SNP 6.0 protocol.

The number of possible matches, given permutation of sequence in loops is intractably large, as is a calculation of the thermodynamic properties of the set of all possible 33mers sequences with all possible variations of both mismatch and insertions. This step was used to guide the development of the alignment filter patterns.
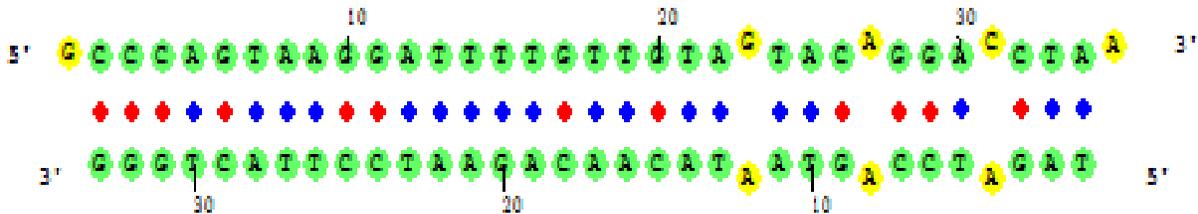

Fig.1 Mismatch:
(ΔG°) @50° C,
-24.72 kcal/mol

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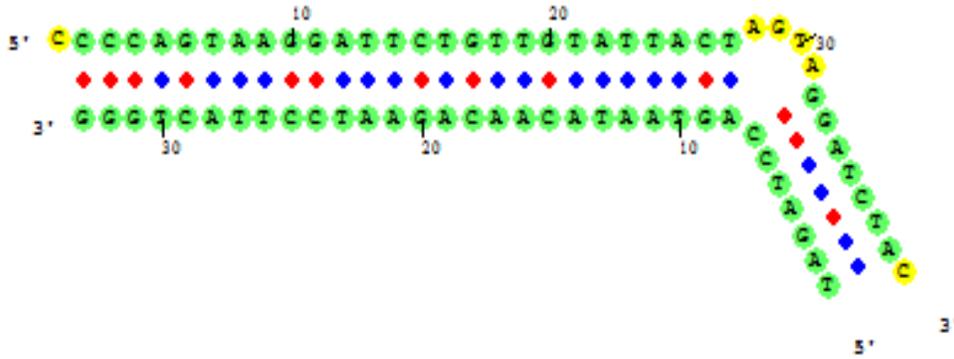(ΔG°) @50° C,
-29.87 kcal/mol

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

Fig. 4 'Bubble':
(ΔG°) @50° C,
-22.01 kcal/mol

Fig. 5 'Tail':
(ΔG°) @50° C,
-22.81 kcal/mol

Fig. 6 'Bubble' and Loop:
(ΔG°) @50° C,
-22.61 kcal/mol

Step 2: Scan Genome for patterns

❖ Pattern Collection: The SeqNFind™ system was employed to generate affine gap alignments for all possible locations within Human RefSeq genome 36.3 that meet our structure-pattern templates. Scoring parameters were adjusted with each iteration.

❖ Pattern Filtering: A tool for filtering the resulting alignment patterns for those that loosely matched with the structures in *step 1* was applied to the resulting sequences to get a library of possible cross hybridizing locations for every probe. The pattern tool was modified with each iteration. Due to the high dependency of ΔG on nearest neighbor parameters, these patterns were designed to allow high amounts of variation.

Step 3: Calculate ΔG

❖ Calculation of Gibbs Free Energy: OMP™ was used in batch mode on the results from the pattern filtering to get the ΔG. The same results were reviewed to look at the variation of ΔG. Parameters for this portion of the run were set to match the hybridization conditions for the Affymetrix SNP 6.0 platform. To simplify the results collated we selected only optimal alignments, discarding all sub-optimal alignments that are returned by OMP. These results guided the patterns and parameters used in *step 2*.

❖ Visual Inspection: A small number of alignment pattern pairs were visually inspected using Visual OMP™; if they did not previously exist in the library they were added and the parameters for *step 2* were adjusted accordingly.

Cross-Hybridization Results

Shown here are results for the 1804 Affymetrix SNP 6.0 probes designed to target SNPs on Chromosome Y, for just the RefSeq positive strand.

Table 1: 1804 probes searched against chrY

Table 1. Column 1 shows the range of ΔG for each row. As is evidenced in Column 2 the number of cross hybridizations goes up as the ΔG increases. What is most startling is the number of perfect match hybridizations found within the [-29, -54] category. Only 44 of the 15306 matches are perfect alignments.

Range ΔG (kcal/mol)	Total Probes	Number of Cross-Hybridizations	Number of Perfect Matches
[-15.0, -22.0)	1804	208685	0
[-22.0, -29.0)	1804	195415	5
[-29.0, -54.0]	1804	68270	536
[-15.0, -54.0]	1804	472370	541

Table 2: 1804 probes searched against ChrY

Table 2. This table shows the amount of variation found on a per probe basis. As in Table 1, Column 1 gives the range for Free Energies. Column 2 give the number of probes that had only 1 optimal score

within the range.   Column 3 shows the maximum number of cross-hybridizations found *for a single probe*. The two blue highlighted entries came *from the same probe* SNP_A-8477444.

Range ΔG (kcal/mol)	Total Probes	# of Probes with only 1 hybridizing location	Maximum # of hybridizing locations for a probe
[-15.0, -22.0)	1804	132	5859
[-22.0, -29.0)	1804	125	8114
[-29.0, -54.0]	1804	50	8572
[-15.0, -54.0]	1804	307	16686

One Probe SNP_A-8280034 versus the Genome

Every probe with a unique sequence will have a unique set of results. As is evident in Table 2, some probes may have a very confounded signal. Conversely, some probes may come back relatively 'clean'. The probe SNP_A-8280034, selected at random to build our library, does not show a large amount of cross hybridization when run against the entire build 36.3 of the genome. Table 3 shows the results of per ΔG category for the single probe. Figures 7 and 8 show two structures discovered for this probe not in our original set of templates (highlighting the iterative process).

Range ΔG	Probe	Number of Cross-Hybridization	Number of Perfect Matches
[-15.0, -22.0)	SNP_A-8280034	128	0
[-22.0, -29.0)	SNP_A-8280034	1	0
[-29.0, -54.0]	SNP_A-8280034	0	0
[-15.0, -54.0]	SNP_A-8280034	129	0

Fig. 7 (up):'Bubbles': (ΔG°) @50º C, -20.44 kcal/mol

Fig. 8 (right):'Bubbles': (ΔG°) @50º C, -23.66kcal/mol

Conclusions and Future Work

From the results above we can conclude that this *in silico* sequence-structure discovery method can be used to find more complete sets of cross hybridizing targets for a given probe. Not all probes behave in the same manner as is seen in Figures 7 and 8. The simple alignment pattern of short indels, seen Fig 7, *may* be found by BLAST-based algorithms. Where there are larger indels those methods will fail; a tool flexible enough to find large indels is required to identify structures like that in Fig 8. A second pass for stability is needed in both cases, to remove the unstable variants from the pool (data not shown). does not bind as well as a structure with a greater number of mismatches..

Acknowledgments

We wish to thank Don Kolva at ATL for his determination, insights and help with the SeqNFind™ system.

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 and 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.

For more information on SeqNFind™, please call or e-mail Accelerated Technology Laboratories, Inc. at info@atlab.com or 800.565.5467 (toll free in the US or Canada) or 910.673.8165.