Genome analysis

Thermodynamic properties of DNA sequences: characteristic values for the human genome

Ryan T. Koehler* and Nicolas Peyret

Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA

Received on April 6, 2005; revised on May 13, 2005; accepted on June 6, 2005

ABSTRACT

Motivation: Central to many molecular biology techniques as ubiquituous as PCR and Southern blotting is the design of oligonucleotide (oligo) probes and/or primers possessing specific thermodynamic properties. Here, we used validated theoretical methods to generate distributions of predicted thermodynamic properties for DNA oligos of various lengths. These distributions facilitate immediate appreciation of typical thermodynamic values for oligos of various lengths.

Results: Distributions of melting temperature ($T_m$), free energy ($\Delta G^\circ_T$), and fraction hybridized or fraction bound (Fb), are presented for oligos of length 10–50 bases sampled from the human genome. The effects of changing temperature, oligo and salt concentrations, constraining G+C content, and introducing mismatches are exemplified. Our results provide the first survey of typical and limiting thermodynamic values evaluated on a genomic scale. Described numbers comprise useful ‘rules of thumb’ that are applicable to most technologies dependent upon DNA oligo design.

Contact: Koehlert@appliedbiosystems.com

Supplementary information: Supplementary Data is available at http://bioinformatics.oxfordjournals.org

1 INTRODUCTION

Molecular biology techniques such as PCR (Saiki et al., 1988), Southern blotting (Southern, 1975), sequencing by hybridization (Fodor et al., 1993), 5’ nuclease assays for allelic discrimination (Livak, 1999), DASH (Prince et al., 2001), molecular beacon based assays (Bonnet et al., 1999), antigen targeting (Freier, 1993) and DNA microarrays (Shoemaker and Linsley, 2002) rely on the great specificity and selectivity afforded by DNA probes. This specificity can be traced to the delicate balance between sequence and thermodynamic properties exhibited by nucleic acids.

Thermodynamic parameters for DNA duplex formation are accurately predicted using the nearest-neighbor model (De Voe and Tinoco, 1962; Gray and Tinoco, 1970; Borer et al., 1974). This model has been extended beyond Watson–Crick sequences to those with mismatches and dangling ends (SantaLucia and Hicks, 2004) and is very accurate for predicting thermodynamic properties of oligonucleotides (oligos) as well as longer nucleotide polymers (SantaLucia, 1998), under a variety of experimental conditions.

An accurate working knowledge of oligo thermodynamic quantities is of great practical importance. Knowledge of melting temperature ($T_m$), free energy of hybridization ($\Delta G^\circ_T$) and fraction of oligos hybridized (Fb) greatly facilitates optimization of oligo sensitivity and specificity for a variety of applications including PCR primer and probe design.

While calculated thermodynamic properties are widely used in oligo design schemes and software (Rychlik and Rhoads, 1989; Peyret and SantaLucia, 1999, http://ozone2.chem.wayne.edu/Hyther/hythermenu.html; Le Novère, 2001), such applications almost invariably focus on achieving specific thermodynamic values for sequences in hand. This narrow scope fails to provide global insight about how important thermodynamic quantities vary in response to changes in sequence, oligo length or experimental conditions. For example, it is essentially impossible to estimate Fb without explicitly calculating it for an oligo sequence. Generally, one cannot even guess upper or lower bounds of thermodynamic values for a generic oligo sequence of some length. Working knowledge of typical and limiting thermodynamic values expected for oligos under a range of experimentally relevant conditions provides a very valuable foundation to anyone undertaking oligo design. For example, one might anticipate that 90% of all possible 16mer DNA oligos completely hybridize under some specific conditions, or that no possible 16mer DNA oligo should hybridize to a target under other conditions. Such rules of thumb can provide a rational starting point for choosing experimental parameters and algorithmic threshold values that are routinely employed in oligo design and screening schemes. Previous studies have focused on the effect of experimental condition variations on random oligo free energy distributions (Rahman and Gräfe, 2004).

Here, we present the first comprehensive survey of oligo thermodynamic quantities evaluated on a genomic scale by studying model oligos sampled from the human genome. Our intent is to provide a generalized overview of typical and limiting values for thermodynamic quantities ($T_m$, $\Delta G^\circ_T$ and Fb) in relation to sequence length (10–50 bases), physical conditions (salt concentration, oligo concentration and temperature), constraints on base composition (G+C composition), nearest-neighbor parameter set [Breslauer (Breslauer et al., 1986) versus SantaLucia (SantaLucia, 1998)] and introduction of mismatches. We have utilized such distributions to guide investigational researchers with assay design at Applied Biosystems for several years (Koehler and Peyret, 2002).

2 METHODS

2.1 Datasets

One hundred thousand (100K) 50 base sequences were sampled from the UCSC human genome (hg17; May 2004; http://hgdownload.cse.ucsc.edu/downloads.html#human), such that sequences were minimally spaced
10,000 bases apart. For each combination of physical conditions and length considered (10–50 bases), subsequences corresponding to the first n bases of the sampled sequences were evaluated using the nearest-neighbor model. In the case where composition was constrained, n-mers with G+C content outside the limited range 25–75% (in the first n bases) were removed prior to evaluation (though 100K samples were still evaluated).

2.2 Calculation of thermodynamic quantities

Calculation of enthalpy ($\Delta H^\circ$), free energy ($\Delta G^\circ$), entropy ($\Delta S^\circ$), $T_m$ and Fb for base pairing of each oligonucleotide to its complement are based on the nearest-neighbor model. The model assumes a two-state cooperative transition between random coils and duplex (Petersheim and Turner, 1983; Freier et al., 1986). The most accurate DNA nearest-neighbor parameters were used for our calculations (Allawi and SantaLucia, 1997, 1998a,b,c; SantaLucia, 1998; Peyret et al., 1999; Bommarito et al., 2000). Breslauer nearest-neighbor parameters (Breslauer et al., 1986) were also used for comparison purposes. Free energy values were corrected for monovalent cation (SantaLucia, 1996) and magnesium concentrations (Peyret, 2000). Detailed equations are provided as Supplementary Material. Minimum folding energy of the strands was calculated to evaluate how secondary structure affects distribution plots. This was done using the Vienna package (Hofacker et al., 1994) parameterized for DNA, using the most up to date parameter set available (SantaLucia and Hicks, 2004).

2.3 Distribution plots

Distributions were generated by evaluating thermodynamic quantities for each of the (100K) randomly chosen n-mers at each length considered. For each length, median and percentile-bounding values were identified and plotted. The final distribution plot is obtained by combining the values obtained for each oligo length. Figure 1 illustrates the construction of a distribution, in this case for $T_m$. Dedicated scripts were used to process and plot (Grubb, 2002) each distribution. Using 100K oligos, distributions obtained between different random samplings of the genome are indistinguishable, indicating that our sampling is robust.

3 RESULTS

As a convenient and useful point of reference, we start by describing thermodynamic quantity distributions ($T_m$, $\Delta G^\circ$ and Fb) calculated with physical conditions routinely used for PCR (Sorschier, 1997). Specifically, we use 50°C temperature, 0.05 M [Na+], 0.0013 M [Mg$^{2+}$], 5 x $10^{-7}$ M [DNA$_A$] and 1 x $10^{-12}$ M [DNA$_{AG}$]. For these reference conditions, DNA$_A$ represents an oligo, such as a PCR primer and DNA$_{AG}$ represents the corresponding target, such as genomic DNA. With distributions for reference conditions in hand, we next explore how introducing dangling ends, perturbing salt concentration oligo concentration and temperature alter oligo thermodynamic values. Differences stemming from alternative nearest-neighbor parameter sets and the impact of restricting oligo G+C content are also discussed. Finally, we study the effect of introducing base mismatches in both non-competing and competing reaction contexts.

Distributions are presented for each thermodynamic quantity evaluated as a function of sequence length. This format allows one to easily identify typical and limiting values for oligos of a given length, as well as to compare distributions of these quantities among lengths. For all plots, a thick black line is used to denote median values and different shades of gray demarcate the center-most 50, 80 and 98 percentiles as illustrated in Figure 1. Tables of all percentile values and additional plots are available as Supplementary Material.

4 DISCUSSION

4.1 Hybridization indicators

Melting temperature is probably the most widely used indicator of duplex stability. If we assume a two-state transition (single-stranded state and duplex state), the melting temperature is the temperature at which half of the limiting nucleic acid strand in solution is single-stranded. A major drawback of $T_m$ as an indicator of duplex stability is the inability to differentiate between oligos that have similar $T_m$ values, when assays are performed at temperatures removed from $T_m$. This is because the fraction hybridized (see below) curve representing the transition from single-strand to duplex may have different slopes as a function of temperature for sequences with similar $T_m$. Another indicator of duplex stability is the free energy of binding, $\Delta G^\circ$. This is sensitive to assay temperature but does not reflect the influence of concentrations in duplex stability. A third indicator that is sensitive to both concentration and temperature is the fraction hybridized, Fb. This expresses the percent of the limiting nucleic acid strand that is in duplex state under given temperature and strand concentration conditions. An Fb of 100% indicates complete hybridization, while an Fb close to zero indicates no appreciable hybridization. $T_m$ is defined as the temperature at which Fb equals 50%. The Fb indicator is clearly the most informative as it directly relates to measurable assay characteristics, such as fluorescence, for example.

4.2 Accuracy of the thermodynamic predictions

SantaLucia (1998) predicted the thermodynamics of polymers that had been experimentally measured and showed that the nearest-neighbor model accurately predicted both oligo and polymer thermodynamics. Consequently, oligos considered in this work (10–50 bases) should be predicted accurately, provided the two-state model (Petersheim and Turner, 1983; Freier et al., 1986) applies. Typical errors in thermodynamic predictions of free energy and $T_m$ are 4% and 2–3°C, respectively (SantaLucia et al., 1996; Allawi and SantaLucia, 1997).

Oligos that have propensity to form secondary structures or slipped duplexes sometimes do not exhibit two-state transitions (Allawi and SantaLucia, 1997). To insure that our distributions are accurate we also evaluated thermodynamics using correction for single-strand folding. For this analysis three coupled equilibria were considered and Fb distributions were deduced (see Supplementary Material for equations). The distributions obtained with and without correction for single-strand folding were essentially identical. This result does not mean that secondary structure should be neglected. However, for the oligos and conditions considered here, significant secondary structure able to disrupt hybridization is rare.
In the case of oligos targeting genomic DNA it is harder to predict potential folding of the target region. For this reason some predictions carry inaccuracies that only bench-work (UV spectroscopy, fluorescent melts, etc.) can resolve. However, collecting experimental data on all potential sequences possible for targeting a genomic region is prohibitively expensive. Computational predictions are, therefore, the main alternative and a required first step in any probe/primer design workflow.

4.3 Reference conditions

Reference distributions of $T_m$, $\Delta G^G_T$ and Fb for oligos evaluated with typical PCR conditions are shown in Figure 2A, B and C, respectively. For these conditions, the median $T_m$ value (Fig. 2A) varies between 32 (10mer) and 74°C (50mer), tending to plateau and increase almost linearly for oligos longer than 30 nt with a slope of $\sim$0.35°C per added base. The width of the $T_m$ distribution spans 35–45°C, though most values (80%) fall within $\sim$10°C of the median for any given length. While typical values are important, consideration of limiting values for any given length is also informative. For example, the plot suggests that virtually no 15mer should have a $T_m$ > 72°C, that no 25mer should have a $T_m$ < 45°C, and that to achieve a $T_m$ of 80°C, an oligo must have at least $\sim$21 bases.

The free energy distribution for our reference conditions (Fig. 2B) shows that median $\Delta G^G_T$ values decrease almost linearly from $-5$ (10mer) to $-37$ kcal/mol (50mer) with a slope of $\sim$0.8 kcal/mol per added base. As length increases, the width of the distribution also increases due to the growing sequence diversity.

The Fb distribution (Fig. 2C) shows that median values range from 0 (10mer) to 100% (10mer and longer sequences). This means that >50% of allsampled 10mers are single-strand, while >50% ofsampled 16mers are fully hybridized at reference conditions. The rather sharp transition from single-strand to fully hybridized occurs over a length corresponding to the addition of 10 bases. The width of the distribution also grows with sequence length to reflect the increasing diversity of the sequence composition. If we only consider the most typical 98% of oligos, we see that <2% of 10mer targets are >50% hybridized, whereas 98% of 32mer targets are essentially 100% hybridized.

In many applications, when hybridizing to genomic DNA segments, probes and primers are shorter than their targets. As a consequence the duplex formed has overhangs. Thermodynamic studies have shown that the first overhanging base is responsible for most of the energetic contribution of the overhang (Riccelli et al., 2002). The stability of DNA sequences with dangling ends has been systematically investigated (Bommarito et al., 2000), and most dangling ends have stabilizing contributions. Comparing distributions with (Fig. 2D–F) and without (Figure S1) dangling ends on both sides of the duplex in reference conditions shows how the distributions shift. Median $T_m$ values shift from 2.6°C (10mers) to $<0.1°C$ (46mers and longer oligos), while the distribution width remains constant. Similarly, lengths required for median Fb values of 100% (full hybridization) are 19 and 20 for sequences with and without dangling ends, respectively. Median free energy values also reflect a 0.4 kcal/mol stabilization when dangling ends are present. The widths of the distributions with and without dangling ends are similar.

4.4 Variation of salt concentration, oligo concentration and temperature

Figure 3A, D and B depict Fb distributions in effective cation concentrations of $1.69 \times 10^{-2}$, $1.69 \times 10^{-1}$ and 1.69 M, respectively. The value $1.69 \times 10^{-1}$ M is derived from the salt correction equation (Santalucia, 1998; Peyret, 2000) (see Supplementary Material, Equation S3) and corresponds to reference conditions with 0.05 M [Na$^+$] and 0.0013 M [Mg$^{2+}$]. As we increase cation concentrations, the higher salt reduces the oligo length required for full hybridization. Looking at the median Fb curves, these lengths are 30, 20 and 15 bases from low to high salt, respectively. Similarly, the $T_m$ distribution curves shift by $\sim$12°C for each 10-fold increment in salt concentration (Figure S2). The distribution shift seen with changes in salt is slightly greater for longer oligos, because of the length dependence present in the salt correction equation.

Figure 3C–E depict Fb distributions with oligo concentration of $5 \times 10^{-9}$, $5 \times 10^{-7}$, and $5 \times 10^{-5}$ M, respectively. The Fb curves show that higher concentrations allow shorter oligos to reach complete hybridization. For instance, median lengths for complete hybridization are 24, 20 and 16 bases for the low, reference and high oligo concentrations, respectively. Similarly, $T_m$ distribution curves (Figure S3) shift by $>12°C$ (10mer) to 3°C (50mer) for each 100-fold increment. For long oligos, the concentration change has less impact as can be inferred from the $T_m$ equation (see Supplementary Material, Equation S5).

Figure 3F, D, and H depict Fb distributions at temperatures of 40, 50 and 60°C, respectively. Higher temperatures require longer lengths for oligos to fully hybridize as evidenced by median Fb, lengths of 15, 20 and 28 bases that are required for full hybridization.
at 40, 50 and 60°C. As expected, this trend is also seen in the free energy distribution which decreases linearly with length. For temperatures of 40, 50 and 60°C, the slope roughly decreases by 0.2 kcal/mol for each 10°C variation (Figure S4).

4.5 Thermodynamic parameter sets

The nearest-neighbor thermodynamic parameters of Breslauer *et al.* (1986) continue to be widely used and implemented in commercial oligo design software, despite the fact that a newer set of unified parameters provide significantly improved predictions (SantaLucia, 1998).

Comparison of predictions performed with each set show that the Breslauer parameters predict shorter oligos to be more hybridized than the unified parameters. For example, the median Fb for full hybridization is 17 bases for Breslauer but 20 bases, using the unified parameters (Figure S5). In accordance with the Fb differences, median \( T_m \) values for 10mers are \( -30°C \) with both parameter sets, but \( T_m \) values for 50mers differ by \( -16°C \) (Figure S5). These observations suggest that the Breslauer parameter set tends to overestimate hybridization stability, and so oligos designed to match a particular \( T_m \) or hybridization level using these parameters may actually be somewhat shorter than intended.

4.6 Constraining sequence composition

Constraining sequence base composition to a G+C of 25–75% for our reference conditions does not alter \( T_m \) and Fb median values (Figure S6). However, the widths of the distributions are significantly reduced when this simple composition filter is applied to limit oligo diversity.

For \( T_m \), the unconstrained width is \( \sim 38°C \) for all considered lengths, but this reduces to \( \sim 25°C \) when the G+C constraint is applied. While the overall distribution width narrows, the width associated with the more typical \( T_m \) values remains largely unchanged. This can be attributed to the fact that most sampled oligos are balanced in terms of G+C content. Sequences that are filtered out have highly skewed G+C, and since \( T_m \) roughly correlates with G+C content, removing these sequences shrinks the distribution width by eliminating extreme cases.
Fig. 4. Effect of single mismatch introduction on $T_m$, and fraction hybridized distributions. (A and B) $T_m$ distributions for perfect match sequences and sequences with one mismatch, respectively are shown. (C) The difference between the two previous $T_m$ distributions is shown. (D and E) Fraction hybridized distributions for perfect match sequences and sequences with one mismatch, respectively are shown. (F) The difference between the two previous fraction hybridized distributions is shown. Experimental conditions correspond to the reference conditions described in the Methods section.

The effect of the G+C constraint on Fb is analogous to the $T_m$ results. Oligos giving rise to extreme values are removed from the sample while the range of more typical values is largely unchanged. From a practical point of view, the base composition example illustrates how application of even simple filters can significantly reduce the variation of oligo thermodynamics.

### 4.7 Sequences with mismatches

The effect of single mismatches on $T_m$, and Fb is shown in Figure 4. Comparing $T_m$ distributions for perfect match and mismatch cases (Fig. 4A and B, respectively) reveals that median $T_m$ values differ from $-20^\circ$C for 10mers to $-2^\circ$C for oligos longer than $\sim$35 bases. This trend is expected, because the relative destabilizing impact of a single mismatch diminishes as oligo length is increased. The distribution of differences between perfect match and mismatch cases is shown in Figure 4C, where the diminishing impact of a single mismatch as length is increased is obvious. Considering limiting cases, we see the most destabilizing single mismatch in a 10 base oligo drops $T_m$ by $>40^\circ$C, and the least destabilizing mismatch drops $T_m$ by $\sim7^\circ$C. For 25mers, the largest drop associated with a single base mismatch is $<10^\circ$C. It is interesting to compare calculated $T_m$ differences to $T_m$ differences obtained in DASH experiments (Prince et al., 2001), where differential melting profiles are used to identify mismatched sequences. For oligos of 15 and 17 bases, $T_m$ differences associated with a single mismatch were found to be $4-12^\circ$C in a magnesium-free buffer. In Figure 4C, the majority (80%) of oligos of this length have calculated $T_m$ differences of $\sim5$ to $\sim14.5^\circ$C with our reference conditions. DASH experiments have also found that increasing probe length from 13 to 25 bases results in a 3-fold decrease in the $T_m$ drop associated with a single mismatch (Howell et al., 1999). The median calculated $T_m$ difference associated with a single mismatch in Figure 4C is $\sim12^\circ$C for 13mers and $\sim4^\circ$C for 25mers, which agrees with the DASH results. Compared to the perfect case, introducing a single mismatch costs the equivalent of $\sim4$ bases on average, if complete hybridization is desired (right shift of the distribution in Fig. 4E versus D). Considering limiting lengths, most sampled 10mer with a single mismatch are $<10\%$ hybridized, and a small number of oligos longer than 40 bases containing a single mismatch is marginally destabilized (i.e. Fb $<100\%$). At reference conditions, the impact of a single mismatch on Fb is most pronounced for oligos of $\sim15-18$ bases, with the largest average Fb difference seen for 17mers. This is evidenced by the hump in Figure 4F. Shorter oligos are, on average, less hybridized, to start with, so destabilizing mismatches cause less disruption than is the case with 17mers. Conversely, longer oligos tend to fully hybridize even with sequence mismatches. Thus under reference conditions, 17mers are most likely to show the greatest discrimination between perfect and mismatch cases. Significantly, consideration of $T_m$ differences alone does not lead one to this result.

We also investigated the effect of multiple base mismatches on hybridization. Fb distributions shift progressively rightward as the number of mismatches increases (Figure S7). This is expected, since longer oligos are required to compensate for the destabilization introduced by the mismatches. The median oligo lengths for full hybridization are $\sim20$, 24, 27 and 35 bases when there are 0, 1, 2 and 4 mismatches, respectively. The mismatch examples considered represent cases where oligos are in excess relative to targets. This situation occurs, for example, when primers or probes are used to amplify and/or interrogate genomic DNA. Any non-target site in the genomic sequence that is similar enough to the intended target (e.g. similar length with few mismatches) may give rise to spurious product and/or readout that compromises an assay. This situation
also occurs with microarrays, where oligos are surface-bound, and each may potentially bind to different target sites in the milieu of interrogated DNA. Non-specific signals will be observed, if conditions allow for binding of non-target sequences with mismatches. We note, however, that since the thermodynamic parameters and equilibrium assumption implicit in the currently used model may not hold for surface-bound oligos (Bahnot et al., 2003), results presented in Figure S7 may not be directly applicable to microarray systems.

An important design consideration for oligos employed in genomic-context assays is to ensure target specificity. This commonly involves scanning the target genome for similar non-target sites, for example, with BLAST (Altschul et al., 1990). Given a list of similar non-target 'hits' to consider, the question then becomes: At what level of similarity to the intended target does a non-target hit pose a risk? Assuming an oligo must appreciably hybridize a non-target site to yield spurious signals, one can identify approximate limits from the mismatch distributions. For example, under reference conditions, most 23mers (90 percentile) fully hybridize to perfect targets while very few hybridize to target-similar sites with four mismatches. Thus (in this system), hits less than ~80% similar to target (e.g. 20/23 matching bases) are unlikely to pose a threat. While specific conditions, assay-imposed design constraints, and even different genomes will probably influence results, theoretical limits of this sort are nonetheless useful starting points for setting algorithmic thresholds.

Figure 5 shows Fb distributions of perfect match oligos that are in competition with single mismatch oligos for a common target strand. Competition in this context means that the target site to which the oligos are hybridizing is at a limiting concentration (i.e. excess oligos) hence either perfect match or mismatch oligos may bind to the target, but not both oligos at once. At equal concentrations, the perfect match oligo overwhelmingly out-competes the single mismatches under reference conditions (Fig. 5A). However, an appreciable fraction of single mismatch oligos longer than ~25 bases is still hybridized to a noticeable extent. This is implied by failure of perfect oligos to achieve Fb = 100%, which yields the light gray band across the top of Figure 5A. In this competition case, total Fb is roughly the sum of perfect and mismatch Fb contributions and has a maximum of 100%. Thus, 1% (99th percentile) of single mismatch oligos have an Fb ~10% (100% minus the 99th percentile trace in Figure 5A) and 10% (90th percentile) have an Fb ~4%. These trends are constant for oligos longer than ~25 bases because both perfect and mismatch sequences favor full hybridization at these lengths, and the free energy ‘cost’ of a single mismatch is independent of length. It is interesting to compare mismatch hybridization with and without competition from the perfect match oligo. With competition from a perfect match oligo, single mismatch oligo hybridization is very limited regardless of length. The fact that most targets for oligos >30 nt are >90% hybridized to perfect match oligos means that these targets cannot be >10% hybridized to mismatched oligos (color band at top of Fig. 5A). Without competition, however, targets for oligos longer than 30 fully hybridize to mismatch oligos (Fig. 5F).

At increased temperature, the free-energy advantage of a perfect match over a mismatch oligo decreases. Because of this, mismatch oligos now viable compete for target sites; this dramatically reduces the number of perfect match oligos that dominantly hybridize. Figure 5B shows the Fb distribution for the perfect match oligo evaluated with the competitive model at 65°C (with equal perfect and mismatch oligo concentrations). The distribution is significantly depressed and right-shifted relative to the 50°C example (Fig. 5A). Distributions of the mismatch Fb (data not shown) reveal that >90% of targets for oligos longer than ~25 bases are >5% hybridized to mismatch oligos at 65°C, while at 50°C ~10% of such targets are hybridized to a similar extent. Thus, increasing temperature decreases relative specificity between perfect and single mismatch oligos when the two oligos compete for the same target site. As expected, increasing the concentration of a mismatch oligo tends to drive the perfect match oligo off the target. This is shown in Figure 5C, where the mismatch oligo is 100-fold in excess of the perfect match oligo under reference conditions. In this case, about half (i.e. median) of the targets for perfect match oligos longer than 25 bases are ~50% bound, 10% (90th percentile) are only ~15% hybridized and 99.9% of targets for perfect match oligos are less than ~97% hybridized.

The above simulation is applicable to single nucleotide polymorphism (SNP) genotyping assays, where two oligos, designed to complement one or the other allelic base at the SNP site, compete for a single target. Assay specificity depends on the perfect match probe largely out-competing the single mismatch probe. For example, 5’ nuclease assays (Livak, 1999), or molecular beacon-based SNP genotyping assays (Marras et al., 1999), require optimized oligo designs so that the mismatched oligo does not generate sufficient signal to make assay readout ambiguous. To further enhance discrimination, modified oligos (i.e. minor groove binder, MGB probes) may also be used (Kutyakin et al., 2000). This simulation is also applicable to microarrays where two surface-bound oligos, differing by a single base, are directed to a given target (Lipshtutz et al., 1999). One oligo matches the target perfectly, while the other contains a mismatch and
is used to evaluate non-specific background signal. A target binding to both perfect and single mismatch oligos lowers the assay signal-to-noise ratio. As noted above, however, the currently used model may not hold for surface-bound oligos so these results are not directly applicable to microarray systems (Bahnot et al., 2003).

5 CONCLUSION

Distribution plots provide a useful starting point for applications requiring design of probes or primers to be employed under various experimental conditions. In addition, the trends made evident as conditions are perturbed provide general insights about the relationship between oligo thermodynamics and experimentally controllable parameters. For example, if oligo length is free to vary while full hybridization is required (e.g. PCR primer design), one can identify upper and lower bounds on the range of oligo lengths likely to be encountered. This can be useful for evaluating potential specificity or synthesis issues (i.e. short oligos are more likely to occur multiple times in a given genome and synthesis of some longer oligos may be problematic). Conversely in situations where length is fixed (e.g. combinatorial probe libraries) conditions may be chosen so that, for example, at least 90% oligos are fully hybridized.

Our goal is to provide a broad overview of the theoretical behavior of DNA thermodynamics from which rules of thumb may be gleaned. Our distribution plots facilitate this by concisely conveying typical and limiting thermodynamic values for oligos of practical length. To our knowledge, this is the first comprehensive analysis of thermodynamic distributions on a genomic scale. We anticipate that distributions presented here should provide useful reference information for researchers devising probe design and screening methodologies. Of course, distribution plots based on actual experimental conditions and assay-imposed oligo constraints will be most appropriate for specific applications.

ACKNOWLEDGEMENTS

We thank Francisco De La Vega, Gene Spier and Hadar Avi-Izhak for helpful suggestions and critical reading of this manuscript.

Conflict of Interest: none declared.

REFERENCES


Peyret,N. and SantaLucia,J., Jr (1999) HYTHER.


