DNA Sequence Context and Multiplex Hybridization Reactions: Melting Studies of Heteromorphic Duplex DNA Complexes

Peter V. Riccelli, Timothy S. Hall, Petr Pancoska, Kathleen E. Mandell, and Albert S. Benight*

Contribution from the Department of Chemistry, University of Illinois at Chicago, 845 West Taylor Street, Room 4500, Chicago, Illinois 60617

Received July 31, 2002; E-mail: abenight@uic.edu

Abstract: Heteromorphic hybrid duplex DNA complexes are duplex states, other than perfectly matched duplexes, that can form when single strands comprising several different perfectly matched duplexes are simultaneously present in solution. Such cross-hybridization "side reactions" are of particular nuisance in multiplex reaction schemes, where many strands are designed to hybridize in parallel fashion with only their corresponding perfect complement strand. Relative to the perfect match duplexes, the sequence dependent features of these heteromorphic duplex states and their thermodynamic stability are an important consideration for multiplex hybridization reaction design. We have measured absorbance versus temperature melting curves and performed differential scanning calorimetry measurements on various mixtures of eight different 24 base single strands. When perfect complementary pairs of strands are mixed in single reactions, four perfectly matched duplexes form. When mixtures of strands that are not perfectly matched are prepared and analyzed, melting transitions for cross-hybridization are observed along with significant hyperchromicity changes. This is indicative of a melting hybrid, heteromorphic duplex states formed from two nonperfectly matched strands. In addition, when both the perfectly matched and noncomplementary strands are mixed together (in multiplex hybridization reactions) at molar ratios of 1:1, 3:1, and 1:3, evidence of perfect duplex and heteromorphic duplex complexes is found in all cases. A new analytical tool for considering heterogeneous, duplex complexes in multiplex hybridization mixtures is presented and employed to interpret the acquired melting data.

Introduction

Large-scale, high throughput, combinatorial approaches to nucleic acid analysis are emerging as powerful tools for widespread applications in detecting, discriminating and analyzing large numbers of DNA sequences via multiplex hybridization schemes.[1-6] By the utilization of these schemes, microarray technologies offer the benefit of being both high throughput and exquisitely accurate.[7-17] This is provided sequences on the microarrays are appropriately designed so desired hybridization reactions occur in the manner intended with high fidelity.

Many microarray platforms utilize solid-support bound oligonucleotide probes to hybridize and thereby capture single strand targets. The majority of formats commonly employ linear single stranded oligonucleotide probes on two-dimensional surfaces (glass slides,[18,19] microtiter plates,[20] gel pads[21]), but bead based formats are also emerging.[22] Alternative probes and probe formats are also being employed.[23,24] Regardless of the

(1) Duggan, D. J.; Bittner, M.; Chen, Y.; Meltzer, P.; Trent, J. M. *Nat. Genet.* **1999**, *21*, 10−14.
(2) Lipshutz, R. J.; Morris, D.; Chee, M.; Hubbell, E.; Kozal, M. J.; Shah, N.; Shen, N.; Yang, R.; Fodor, S. P. *BioTechniques* **1995**, *19*, 442−447.
(3) O'Donnell-Maloney, M. J.; Smith, C. L.; Cantor, C. R. *Trends Biotechnol.* **1996**, *14*, 401−407.
(4) de Saizieu, A.; Certa, U.; Warrington, J.; Gray, C.; Keck, W.; Mous, J. *Nat. Biotechnol.* **1998**, *16*, 45−48.
(5) Southern, E. M.; Mir, K.; Shchepinov, M. *Nat. Genet.* **1999**, *21*, 5−9.
(6) Chen, J. J.; Wu, R.; Yang, P. C.; Huang, J. Y.; Sher, Y. P.; Han, M. H.; Kao, W. C.; Lee, P. J.; Chiu, T. F.; Chang, F.; Chu, Y. W.; Wu, C. W.; Peck, K. *Genomics* **1996**, *51*, 313−324.
(7) Cheung, V. G.; Morley, M.; Aguilar, F.; Massimi, A.; Kucherlapati, R.; Childs, G. *Nat. Genet.* **1999**, *21*, 15−19.
(8) Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. *Science* **1995**, *270*, 467−470.
(9) Ramsay, G. *Nat. Biotechnol.* **1998**, *16*, 40−44.
(10) Shalon, D.; Smith, S. J.; Brown, P. O. *Genome Res.* **1996**, *6*, 639−645.
(11) Khan, J.; Saal, L. H.; Bittner, M. L.; Chen, Y.; Trent, J. M.; Meltzer, P. S. *Electrophoresis* **1999**, *20*, 223−229.

(12) Szallasi, Z. *Pac. Symp. Biocomput. '99* **1999**, 5−16.
(13) Chee, M.; Yang, R.; Hubbell, E.; Berno, A.; Huang, X. C.; Stern, D.; Winkler, J.; Lockhart, D. J.; Morris, M. S.; Fodor, S. P. *Science* **1996**, *274*, 610−614.
(14) Pease, A. C.; Solas, D.; Sullivan, E. J.; Cronin, M. T.; Holmes, C. P.; Fodor, S. P. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 5022−5026.
(15) Fodor, S. P.; Rava, R. P.; Huang, X. C.; Pease, A. C.; Holmes, C. P.; Adams, C. L. *Nature* **1993**, *364*, 555−556.
(16) Shoemaker, D. D.; Lashkari, D. A.; Morris, D.; Mittmann, M.; Davis, R. W. *Nat. Genet.* **1996**, *14*, 450−456.
(17) Eisen, M. B.; Spellman, P. T.; Brown, P. O., Botstein, D. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14863−14868.
(18) Shalon, D.; Smith, S. J.; Brown, P. O. *Genome Res.* **1996**, *7*, 39−45.
(19) Kane, M. D.; Jatkoe, T. A.; Stumpf, C. R.; Lu, J.; Thomas, J. D.; Madore, S. J. *Nucleic Acids Res.* **2000**, *22*, 4552−4557.
(20) Green, E. D.; Olson, M. V. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *3*, 1213−1217.
(21) Tillib, S. V.; Strizhkov, B. N.; Mirzabekov, A. D. *Anal. Biochem.* **2001**, *1*, 155−160.
(22) Spiro, A.; Lowe, M.; Brown, D. *Appl. Environ. Microbiol.* **2000**, *10*, 4258−4265.
(23) Westin, L.; Miller, C.; Vollmer, D.; Canter, D.; Radtkey, R.; Nerenberg, M.; O'Connell, J. P. *J. Clin. Microbiol.* **2001**, *3*, 1097−104.
(24) Riccelli, P. V.; Merante, F.; Leung, K. T.; Bortolin, S.; Zastawny, R. L.; Janeczko, R.; Benight, A. S. *Nucleic Acids Res.* **2001**, *29*, 996−1004.

format, hybridization of nucleic acid targets to tethered oligo-nucleotide probes in multiplex fashion is the central event in the detection of nucleic acids on oligonucleotide microarrays. Effective design based on target sequences and their sequence dependent stability is essential to achieve optimum hybridization performance. Our focus is on optimum design of DNA sequences for multiplex hybridization.

Designing sequences for multiplex reactions requires consideration of two aspects, informatics and engineering. The informatics aspect concerns the process of defining sequences uniquely diagnostic of the desired targets. Much attention has been paid to this important component, and a number of methods tout "better" sequence screening and selection capabilities for identifying unique target sequences.[25–28] In contrast, considerably less attention has been paid to the, perhaps equally important, engineering aspect.

Once uniquely diagnostic sequences from desired targets have been identified, subsets of sequences must be chosen that will hybridize in a highly specific manner (with reliable and predictable precision) with their complementary probes in the presence of other sequences. This is the engineering aspect of sequence design and involves designing sequences that will exhibit predictable and consistent hybridization characteristics such as uniform signal intensity, comparable thermostability, and zero cross-hybridization with other strands. Problems associated with the engineering aspect of sequence design are the subject of this report. In particular, our approach considers consequences of sequence context on cross-hybridization in multiplex hybridization reactions.

Cross-hybridization results when heteromorphic "hybrid" duplex states, other than the perfectly matched duplex, form. This occurs when single strands from several different perfectly matched duplexes are simultaneously present in solution. Obviously, such "side reactions" are of a particular nuisance in multiplex reactions. They result in lower accuracy and decreased precision and sensitivity and complicate interpretations of microarray data. Effective sequence design should minimize the likelihood for such reactions and thereby allow for higher quality data.

Historically, evaluations of DNA sequence dependent thermodynamic stability have focused entirely on studies of the thermodynamic behaviors of solutions of homogeneous populations of individual strands that anneal to form a single type of duplex molecule.[29–31] This is so-called homogeneous hybridization. In contrast, in multiplex or heterogeneous hybridization, many target single strands are present in the same reaction. If their sequences are so predisposed, these target strands can anneal with other (target and probe) strands that are not fully complementary, forming reasonably stable, partially duplex states that have enough favorable interactions to be relatively stable at assay temperatures. Consideration of the stabilizing interactions that give rise to these cross-hybridizing states, and

determining their relative stability and probability of occurrence, is essential for accurate sequence design and selection for multiplex hybridization schemes. Sequence context is a critical determinant of cross-hybridization propensity, but until now, effects of sequence context on cross-hybridization have not been considered. Here, sequence context is defined with regard to heterogeneous mixtures of strands. Such heterogeneous mixtures of strands occur in multiplex reactions, where several or many different duplexes comprised of their constituent single strands are present. Concurrently, context refers not only to the order and composition of base pairs in each perfect duplex but also to the sequences and contexts of all the other strands and their relative complementarity with respect to the perfect duplexes that are present.

As reported here, we have developed a design tool for treating heterogeneous, multiplex hybridization mixtures on the basis of a new representation of DNA sequences. For each sequence, an ensemble of sequence specific configurations that depend explicitly on the identity and context of the entire sequence contributes to the character of a so-called context function. This function is termed the context functional descriptor (CFD). The CFD has a direct physical meaning serving as a means of coding context dependent sequence information in a convenient and useful functional form, which provides new vistas for analytical comparisons of DNA sequences. We also present analysis, via the CFD approach, of a homogeneous hybridization system, where context effects on the stability of a single base pair mismatch were reported.[32]

Materials and Methods

DNA Molecules. Sequences of the eight 24 base DNA strands that were prepared for melting studies are shown in Figure 1. These strands are designated I_P, II_P, III_P, IV_P and I_T, II_T, III_T, IV_T. Eight different pair combinations of these strands were studied. Four pairs, I_P-I_T, II_P-II_T, III_P-III_T, and IV_P-IV_T, are perfectly complementary and serve as the controls, or reference duplexes. The other four pairs of strands that were prepared and studied are II_T-I_P, I_T-II_P, IV_T-III_P, and III_T-IV_P. These combinations represent potential cross-hybridization products when two duplexes are both present in the same hybridization reactions.

DNA strands were purchased from Oligo's etc and received in gel purified form. Extinction coefficients were determined for each strand according to established methods.[33] The extinction coefficient values in units of cm^{-1} M^{-1} per nucleotide for I_P, II_P, III_P, and IV_P are 9228.3, 8426.8, 8749.2, and 9038.2, respectively, and 8656.6, 8549.0, 9293.8, and 8463.1 for I_T, II_T, III_T, and IV_T, respectively. Strand concentrations were determined from absorbance readings at 260 nm using the aforementioned extinction coefficients.

Optical Melting Curves. Absorbance versus temperature curves were collected for the different strand mixtures in buffered solvent containing 85 mM Na^+ (75 mM NaCl, 10 mM sodium phosphate, 0.1 mM EDTA, pH = 7.8). Sample preparation and data collection procedures were essentially those described previously.[34,35] Total strand concentrations for all optical melting experiments were ~2.5 μM. Absorbance measurements of buffered DNA solutions were collected at 268 nm, while temperature was increased at a linear rate (~60 °C/h) from 12 to 73 °C. Displayed curves are plotted as the relative

(25) Lee, M. K.; Lynch, E. D.; King, M. C. *Genome Res.* **1998**, *3*, 306−312.
(26) Zhang, M. Q. *Genome Res.* **1999**, *8*, 681−688.
(27) Buck, G. A.; Fox, J. W.; Gunthorpe, M.; Hager, K. M.; Naeve, C. W.; Pon, R. T.; Adams, P. S.; Rush, J. *BioTechniques* **1999**, *3*, 528−536.
(28) Talaat, A. M.; Hunter, P.; Johnston, S. A. *Nat. Biotechnol.* **2000**, *6*, 679−682.
(29) Benight, A. S.; Pancoska, P.; Owczarzy, R.; Vallone, P. M.; Nesetril, J.; Riccelli, P. V. *Methods Enzymol.* **2001**, *340*, 165−92.
(30) Owczarzy, R.; Vallone, P. M.; Gallo, F. J.; Paner, T. M.; Lane, M. J.; Benight, A. S. *Biopolymers* **1997**, *44*, 217−239.
(31) SantaLucia, J., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1460−1465.

(32) Hall, T. S.; Pancoska, P.; Riccelli, P. V.; Mandell, K. E.; Benight, A. S. *J. Am. Chem. Soc.* **2001**, *123*, 11811−11812.
(33) Henderson, J. T.; Benight, A. S.; Hanlon, S. *Anal. Biochem.* **1992**, *201*, 17−29.
(34) Riccelli, P. V.; Vallone, P. M.; Kashin, I.; Faldasz, B. D.; Benight, A. S. *Biochemistry* **1999**, *38*, 11197−11208.
(35) Vallone, P. M.; Benight, A. S. *Biochemistry* **2000**, *39*, 7835−7846.

I_P

```
5'-GTT ATG ATT GTA GAT AAA AGG ATT-3'
```

I_T

```
5'-AAT CCT TTT ATC TAC AAT CAT AAC-3'
```

II_P

```
5'-AAG AGA TTG TAT TGT AGA TAA AAG-3'
```

II_T

```
5'-CTT TTA TCT ACA ATA CAA TCT CTT-3'
```

III_P

```
5'-GTT AGA TTT GAT GTA TTG TAT TGA-3'
```

III_T

```
5'-TCA ATA CAA TAC ATC AAA TCT AAC-3'
```

IV_P

```
5'-TTG AGT ATG ATT TGT ATG ATA GAA-3'
```

IV_T

```
5'-TTC TAT CAT ACA AAT CAT ACT CAA-3'
```

Figure 1. Sequences of the four target strands (I_T, II_T, III_T, IV_T) and four probe strands (I_P, II_P, III_P, IV_P) that were mixed in different pairs to form the perfectly matched duplex and heteromorphic duplex complexes that were studied. All strands are 24 bases in length. Perfectly matched duplexes were formed by mixing identically numbered target, T, and probe, P, strands; that is, I_P + I_T forms perfect duplex I_P–I_T. Heteromorphic duplexes described in the text formed spontaneously with the mixing of different pairs of probe and target strands.

absorbance, RelA_{268} = $A_{268}(T)/A_{268}(T_1)$ − 1.0, versus temperature, where $A_{268}(T_1)$ is the absorbance at the first temperature.

Differential Scanning Calorimetry. Measurements of the excess heat capacity, ΔC_p versus temperature, were made using a Nano II DSC instrument (Calorimetric Sciences Corp.). Procedures for sample preparation, data collection, and analysis were essentially those described previously.[34,35] Total strand concentrations for DSC experiments ranged from 110 to 150 μM.

Analytical Methods

The analytical approach that we employed is based on using a novel "functional" representation of DNA sequences. This representation is termed the context functional descriptor, CFD. In principle, a CFD can be constructed for every combination of two strands that comprises a single duplex. For example, the duplex comprised of strands A and B has a CFD for all possible sequence pairs in an antiparallel orientation, that is, A–B, A–A, and B–B. Each CFD is constructed by aligning two strands (5′–3′/3′–5′) and sliding one strand over the other, one base at a time, base by base and estimating experimental t_m or thermodynamic parameters ΔG, ΔH and ΔS for the hybrid duplex state at each alignment step. This strand alignment scheme is depicted in Figure 2a. A plot of the estimated parameters (t_m, ΔG, ΔH, ΔS, etc.) for the hybrid duplex state at each alignment step versus the alignment position provides the CFD. The quantitative meaning of each point on the CFD depends on the parameters used. In the current treatment, a number of aligned base pairs with corresponding hydrogen bonding contributions, nearest-neighbor dependent stacking interactions, and parameters for nearest-neighbor dependent single base pair mismatches are considered. The values employed are taken from the literature.[31,36–39]

(36) Owczarzy, R.; Vallone, P. M.; Goldstein, R. F.; Benight, A. S. *Biopolymers* **2000**, *52*, 29–56.
(37) Allawi, H. T.; SantaLucia, J., Jr. *Biochemistry* **1998**, *37*, 9435–9444.
(38) Allawi, H. T.; SantaLucia, J., Jr. *Biochemistry* **1998**, *37*, 2170–2179.
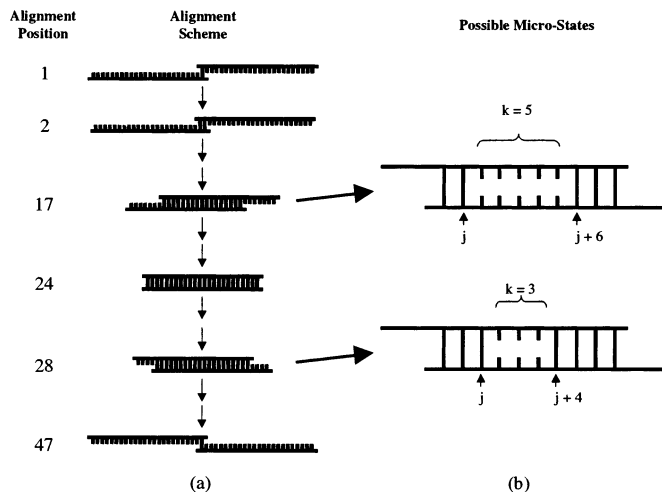(39) Allawi, H. T.; SantaLucia, J., Jr. *Biochemistry* **1997**, *36*, 10581–10594.

Figure 2. Construction of the context functional descriptor (CFD). (a) The sliding algorithm in which a pair of strands is aligned in an antiparallel orientation and one strand is slid across the other, one base at a time. Different possible alignments at the starting point (position 1), positions 2, 17, 24 (complete overlap), 28, and 47 (the last possible alignment) are shown. (b) Hypothetical tandem mismatch states that can form at various alignment positions. For example, at alignment positions 17 and 28, the indicated hypothetical states contain $k = 5$ and $k = 3$ tandem mismatch bases bounded by intact base pairs at positions j and $j + k + 1$. Our treatment of the thermodynamic contributions of such tandem mismatch states to the overall stability of any given heteromorphic complex is described in the text.

In addition, the stability of configurations in states with more than one mismatch in a row (referred to as tandem mismatch states) is estimated from literature values for the sequence dependence of single base pair mismatches. Sequence dependent stabilizing interactions that can occur within such tandem mismatches are also considered in constructing quantitatively meaningful CFDs. As constructed, the CFD of each duplex serves as a semiquantitative functional signature of the relative stability of the ensemble of heteromorphic duplex states that can form between two strands. Thus, the particular shape of the CFD explicitly depends on the sequence identity, composition, and arrangement and should depend on the overall sequence context. Our results support this assertion.

Whatever parameters are used to evaluate the CFD, several standard features of the function should be emphasized. The CFD strongly depends on the actual base ordering in the respective strands. The perfect duplex alignment provides the extreme value (maximum or minimum depending on the convention) on the CFD. In this representation, instead of an N base pair duplex being characterized by a single value, t_m, ΔG, ΔH, or ΔS (or all four values), the N base pair duplex sequence is represented by a function having 2·N−1 points. Using the thermodynamic parameters in this way, the deepest minima and highest maxima on the CFD bracket the most stable fully formed duplex states and least stable partially base paired duplex complexes, respectively.

For the thermodynamic treatment of tandem (two or more) base pair mismatches, which could likely occur in one of the heteromorphic hybrid duplexes depicted in Figure 2a, consider the specific situations depicted in Figure 2b. The hybrid duplex states shown have k mismatches sandwiched between intact pairs at positions j and $j + k + 1$. Examples are for $k = 5$ and $k = 3$. In our analysis, the enthalpic contribution (for example) of this local state (considering only the interactions directly involving mismatches) is given by

$$\Delta H_L = \Delta H_{bp}(j) + \Delta H_{bp}(j+k+1) + \Delta H_{MM}(k) \qquad (1)$$

where $\Delta H_{bp}(j)$ and $\Delta H_{bp}(j+k+1)$ are the enthalpic contributions from the intact base pairs at positions j and $j + k + 1$ and $\Delta H_{MM}(k)$ is the enthalpic contribution from the k tandem mismatches. To estimate the

value of $\Delta H_{MM}(k)$, we use the nearest-neighbor single base pair mismatch parameters as follows:

$$\Delta H_{MM}(k) = 1/k \sum_{i=1}^{k} \Delta H_{j+i-1, j+i, j+i+1} \quad (2)$$

The term in the sum $\Delta H_{j+i-1, j+i, j+i+1}$ is the nearest-neighbor dependent enthalpy for the single base pair mismatch at position $j + 1$ with the specific neighboring base pairs at positions $j + i - 1$ and $j + i + 1$. In essence, we treat each mismatch in the tandem mismatch group as a single base pair mismatch and average over the nearest-neighbor dependent single base pair mismatch values for those sequences comprising the tandem mismatches. It would seem treating tandem mismatches in this manner might overestimate the stability of the tandem mismatch state. This is because (presumably) stabilizing sequence dependent interactions of single base pair mismatches with neighboring intact base pairs should comprise a portion of the nearest-neighbor single base pair mismatch parameters. Actually, intact base pairs do not exist within tandem mismatches; thus, using the single base pair mismatch parameters would be expected to overestimate the stability. However, as will be shown, consideration of the presence of even additional stabilizing sequence dependent interactions within tandem mismatch groups is essential to provide improved agreement with experimental observations for heteromorphic complexes. In fact, these stabilizing interactions add even more stabilizing influence, in addition to that contained in the mismatch pair, to the calculated stability of the tandem mismatch loop. This is in addition to that already obtained from using the single base pair mismatch nearest-neighbor parameters, which is required for best agreement with experiment. Because an additional stabilizing interaction must be included, our conventional estimates on stability contributions from tandem mismatches, as given in eq 2 for example, provide a lower limit estimate on the overall stability of tandem mismatch states. These required additional new considerations reveal that mismatch loops comprising tandem mismatched base pairs are fundamentally different from internal loops consisting of broken base pairs. This difference is explicitly considered in our treatment.

In our approach, the most current sequence dependent stability parameters, evaluated to the highest necessary order or interaction, are utilized. Parameters for nearest-neighbor base pairs and single base pair mismatches bounded by specific nearest-neighbors are utilized to make each point on the CFD as quantitative as possible. Here, we used the sequence dependent parameters inherent in the Hyther program or our recently reported nearest-neighbor parameters.[40,36]

In the nearest-neighbor model, the enthalpy of the duplex, ΔH_{duplex}, is written in terms of the hydrogen bonding component, ΔH_{H-bond}, that depends only on the number of A–T (T–A) and G–C (C–G) and the nearest-neighbor interaction component, ΔH_{n-n},[29] determined according to

$$\Delta H_{duplex} = \Delta H_{H-bond} + \Delta H_{n-n} =$$
$$\Delta S_{bp} [N_{AT}T_{AT} + N_{GC}T_{GC}] + \sum_{ij}N_{ij}(\delta H_{ij}) + \Delta H_M \quad (3)$$

where N_{AT} and N_{GC} are the numbers of A·T or G·C type base pairs in the duplex sequence. The average melting temperatures of A·T or G·C base pairs are given by T_{AT} or T_{GC}.[29] The summed term in eq 3 includes the n-n sequence dependence. N_{ij} is the number of times the n-n doublet ij ($ij = 1-10$) occurs in the duplex sequence, and δH_{ij} is the deviation from the average nearest-neighbor dependent enthalpy for sequence doublet ij.[29] The final term, ΔH_M, accounts for single base pair mismatches or tandem mismatches, such as might occur in certain aligned states other than the perfect duplex. Obviously, for the

(40) SantaLucia, J., Jr.; Peyret, N. HYTHER server and Department of Chemistry, Wayne State University: Detroit, MI. *http://jsl1.chem.wayne.edu/Hyther/hytherobt.html*.

perfect duplex, $\Delta H_M = 0$. For the single base pair, tandem and larger mismatches, $\Delta H_M = \Delta H_{mm} + \Delta H_{MM}$, where ΔH_{mm} represents the nearest-neighbor dependent single base pair mismatch parameters[40] and ΔH_{MM} is calculated for tandem mismatches according to eq 2.

The entropy change of base pair melting, ΔS_{bp}, is assumed to be independent of the sequence. In our calculations, we use the recently reported value for DNA oligomers at -22.4 cal/K mol bp.[31] The total transition entropy of the duplex is simply

$$\Delta S_{duplex} = \Delta S_{bp} [N_{AT} + N_{CG}] \quad (4a)$$

For aligned states with partial duplex overlap and single strand overhangs on the ends,

$$\Delta S_{duplex} = \Delta S_{bp} [N_{bp} + N_{loop}] \quad (4b)$$

where N_{bp} is the number of base pairs in the overlapped duplex regions and N_{loop} is the number of single base pair, tandem base pair, and larger mismatch loops.

The transition temperature, T_m, is calculated according to

$$T_m = (\Delta H_{duplex} + \Delta H_{nuc})\Big/\left(\Delta S_{duplex} + \Delta S_{nuc} + R \ln\left(\frac{[C_T]}{4}\right)\right) \quad (5)$$

where ΔH_{nuc} and ΔS_{nuc} are the enthalpy and entropy of nucleation, respectively. A value of -9.0 cal/K mol for ΔS_{nuc} was employed. The nucleation enthalpy was determined according to,[29]

$$\Delta H_{nuc} = H_1 - (H_2 f) - (H_3 N_{overlap}) \quad (6)$$

where the values of H_1, H_2, and H_3 were 7654.71, 3469.93, and 186.51 and the value of f depends on whether the duplex is a perfect duplex or overlapped duplex. For a perfect duplex, $f = f_{GC}$, the fraction of CG base pairs in the duplex.[29,30] For an overlapped duplex, $f = f_{bp}$, the fraction of intact base pairs in the overlapped region. For a perfect duplex, $N_{overlap}$ is just the total number of base pairs. For other aligned states, $N_{overlap}$ is the number of overlapped bases in each aligned configuration.

Results

Optical Melting Curves. Melting curves of the four perfectly matched duplexes comprised of a probe (P) and target (T) strand (Figure 1) with the designations I_P–I_T and II_P–II_T, formed from a 1:1 mixture of the probe and target strands alone in solution, are shown in Figure 3a. The curves for duplexes III_P–III_T and IV_P–IV_T, formed from a 1:1 mixture of the strands alone in solution, are shown in Figure 3b. These curves are displayed in nearly primary form as plots of the relative absorbance increase at 268 nm, Rel A_{268}, versus temperature. The raw data were smoothed from experimental noise by being passed through an NER filter.[41] Values of the experimentally determined melting temperature, t_m, of each curve, determined as the maximum on temperature derivative curves of Rel A_{268} versus temperature (not shown), are given in Table 1. The t_m values calculated using the HyTher program[40] and our nearest-neighbor parameters[36] (us) are also given in Table 1.

The following hybrid strand combinations, I_P–II_T, II_P–I_T, III_P–IV_T, and IV_P–III_T, were also prepared in 1:1 molar ratios of strands, and their melting curves were measured. Total strand concentrations for these experiments were 2.5 μM (1.25 μM each probe and target strand). The first two pairs are possible combinations if both duplexes I_P–I_T and II_P–II_T are present. A

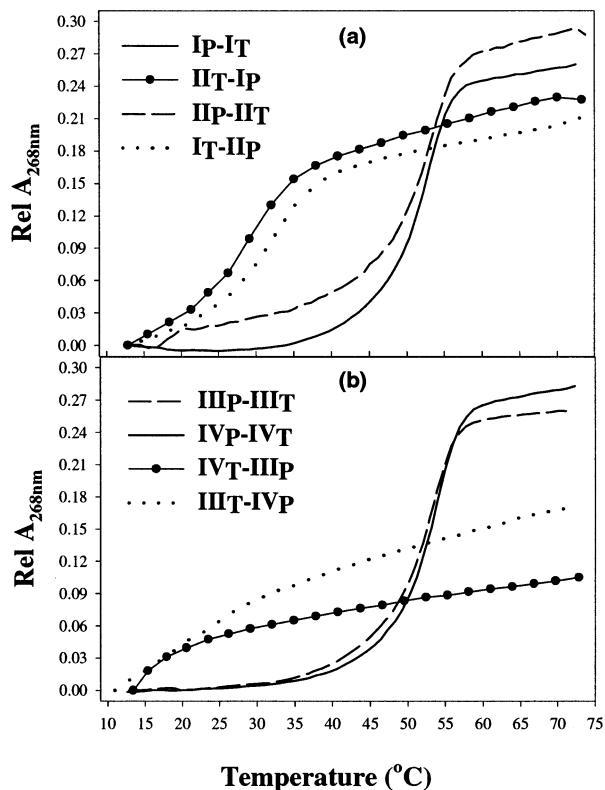(41) Kaiser, J. F.; Reed, W. A. *Rev. Sci. Instrum.* **1977**, *48*, 1447–1457.

Figure 3. Optical melting curves of the perfect match probe−target complexes and four heteromorphic complexes. Plots of relative absorbance measured at 268 nm versus temperature are shown for a 1:1 molar mixture of the different pairs of strands indicated. (a) Melting curves for the perfect match duplexes ($I_T−I_P$ and $II_T−II_P$) and heteromorphic duplexes ($I_T−II_P$ and $II_T−I_P$). (b) Melting curves for the perfect match duplexes ($III_T−III_P$, $IV_T−IV_P$) and heteromorphic duplexes ($III_T−IV_P$, $IV_T−III_P$). Curves are denoted by the keys on the figures. In both cases examined, heteromorphic complexes form which produce significant hyperchromicity with increased temperature. The transition temperature, t_m, values determined from these curves are summarized in Table 1.

Table 1. Experimental and Calculated t_m Values for the Perfect Match Duplexes and Heteromorphic Complexes Studied

	t_m, (°C)			
duplex	experimental[a]	us[b]	HyTher[c]	us[d] (optimized)
$I_T−I_P$	52.7	52.0	51.0	52.0
$II_T−II_P$	52.7	51.5	50.6	51.5
$I_T−II_P$	32.5	9.0	36.1	31.1
$II_T−I_P$	29.2	7.1	36.5	29.2
$III_T−III_P$	51.7	53.3	51.2	53.3
$IV_T−IV_P$	52.6	52.8	51.6	52.8
$III_T−IV_P$	∼15	−23.6	−32.4	3.3
$IV_T−III_P$	∼17	−13.6	2.3	19.5

[a] Experimental uncertainty is ± 1 °C. [b] Calculated using parameters reported in ref 36 and mismatch parameters from ref 40. [c] Values determined by employing HyTher program.[40] [d] Values determined using parameters reported in ref 36, plus stabilizing contribution of tandem mismatch loops.

situation analogous to both being present in the same multiplex hybridization mixture. The latter two pairs are possible combinations when both $III_P−III_T$ and $IV_P−IV_T$ are present. The collected melting curves of the $I_P−II_T$ and $II_P−I_T$ hybrids are shown in Figure 3a, in comparison with the curves of the perfect match duplexes. The melting curves for the $III_P−IV_T$ and $IV_P−III_T$ hybrids are shown in Figure 3b. The t_m values estimated using the HyTher program are given in Table 1. The estimates obtained using our nearest-neighbor parameters[36] in
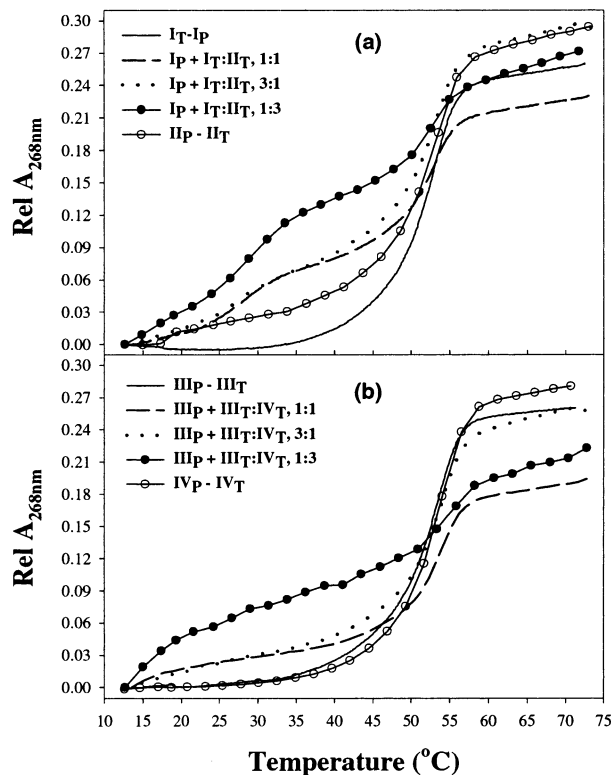

Figure 4. Optical melting curves of mixtures of disparate concentrations of probe and target strands. Relative absorbance measured at 268 nm (Rel A_{268nm}) versus temperature is plotted for mixtures of probe strands and different molar ratios of two different added target strands. (a) Melting curves for the perfect match duplexes ($I_T−I_P$ and $II_T−II_P$) and the heteromorphic duplex ($II_P−I_T$) formed when II_T is mixed at molar ratios of 1:1, 3:1, and 1:3 of the perfect match duplex ($I_T−I_P$). Corresponding curves are denoted by the key on the figures. (b) Melting curves for the perfect match duplexes ($III_T−III_P$, $IV_T−IV_P$) and the heteromorphic duplex ($IV_T−III_P$) formed by adding IV_T at 1:1, 3:1, and 1:3 ratios of the perfect duplex ($III_T−III_P$). In all three multiplex environments, heteromorphic duplex formation occurs to a notable extent, even when competing with perfect duplex formation.

combination with the HyTher^{TM} [40] mismatch numbers (us), and including the calculation for the tandem mismatches, as diagramed in Figure 2b, are also given in Table 1.

As seen on the melting curves in Figure 3, the hyperchromicity changes for the four perfect match duplexes range from 22% to 29%, consistent with what might be expected for the melting of 24 base pair duplexes. As seen by comparison, hyperchromicity changes for the hybrid mixtures were only slightly less for the $I_P−II_T$ and $II_P−I_T$ mixtures (18−22%) but significantly lower for the $III_P−IV_T$ and $IV_P−III_T$ mixtures (9−15%). Albeit relatively less than those observed on melting curves of the perfect match duplexes, these hyperchromicity changes for the hybrid mixtures reveal some amount of heteromorphic complex formation. The t_m values estimated from these hybrid melting curves are also summarized in Table 1 and, as expected, are lower than those for the perfectly matched duplexes.

The melting behaviors of disparate strand mixtures containing the perfect matched duplex strands and a competing strand that can form a hybrid heteromorphic complex with one of the strands of the perfect duplex were also investigated. For reference, the melting curves of the $I_P−I_T$ and $II_P−II_T$ perfect duplexes are displayed in Figure 4a. Total strand concentrations for these experiments were ∼2.5 μM.

Probe concentrations were 1.25 μM, and the total concentration of the perfect and competing cross-hybridizing target strand was 2.5 μM. The melting curves for the perfect duplexes are monophasic with a single melting transition. In contrast, melting curves of the I_P-I_T duplex in an equal molar mixture with the II_T strand and with a 3:1 molar excess of II_T display a distinct pretransition shoulder or subtransition at the same temperature as that of the transition of the I_P-II_T hybrid mixture alone. At an excess molar ratio of the competing II_T strand, the subtransition is a greater fraction of the total hyperchromicity of the overall melting transition. These observations indicate that the perfect duplex and hybrid duplex can coexist if their sequences are conducive for heteromorphic duplex formation. This is precisely a case that could be encountered in a multiplex hybridization reaction, where such strands might be mutually present. A similar observation is made for the melting curves of the III_P-III_T perfect duplex mixed with the third strand IV_T shown in Figure 4b. This subtransition has a relatively lower intensity but increases with increased concentrations of the third strand, and it occurs over the same temperature range as that observed for the melting transition of the III_P-IV_T hybrid mixture alone (Figure 3b). The plots in Figure 4 clearly demonstrate the coexistence of the perfect duplexes and heteromorphic duplexes in a situation analogous to that of a multiplex hybridization reaction.

Optical melting curves were also measured for each of the individual strands (not shown). When the temperature was raised from 20 °C to 90 °C, none of these displayed significant hyperchromicity (less than 3%). These observations indicate the hyperchromicity changes observed on the melting curves for the various strand mixtures must be due to a duplex structure formed when at least two different strands are present. They further suggest that the observed melting transition does not arise from intrastrand structures or interstrand complexes formed from two identical strands.

DSC Melting Curves. Independent confirmation of heteromorphic complex formation for the III_P-IV_T and IV_P-III_T hybrid mixtures was also obtained from differential scanning calorimetry (DSC). The DSC melting transitions for the perfect match duplex IV_P-IV_T and the III_P-IV_T and IV_P-III_T hybrid mixtures alone are shown Figure 5. Total strand concentrations for these experiments ranged from 110 to 150 μM. On the DSC melting curves in Figure 5, the measured excess heat capacity change, ΔC_p, is plotted versus temperature. The melting transitions of the hybrid mixtures are of a considerably lower intensity with t_m values around 20 °C but are clearly measurable. Thermodynamic parameters of the melting transitions evaluated from the DSC measurements are summarized in Table 2. These provide direct evidence of the heteromorphic duplex complexes, with significant ΔH, ΔS, and temperature dependent stability. For the IV_P-III_T hybrid, the transition parameters ΔH and ΔS are approximately 26% of the values measured for the perfect match duplex. For the III_P-IV_T hybrid, the measured thermodynamic parameters are a larger percentage of the perfect match values and the experimentally determined ΔH and ΔS are approximately 40% of the perfect match values. Obviously, both heteromorphic complexes have significant stability. The free-energy values (20 °C) for the hybrid duplexes are slightly negative (−0.6 and −1.1 kcal/mol which are unfavorable in this convention) compared to approximately 20 kcal/mol for the perfect match.
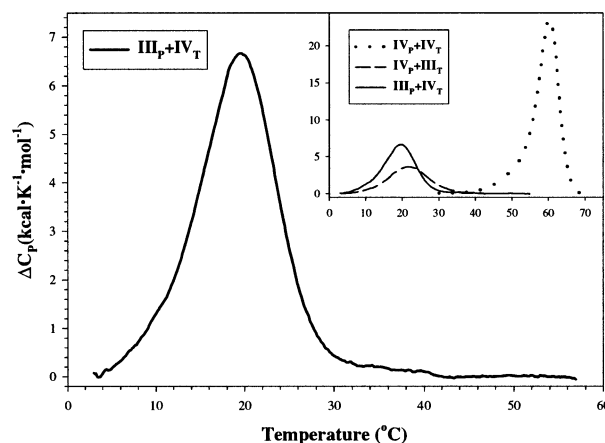

Figure 5. Differential scanning calorimetry (DSC) melting curves for the perfect match duplex and two hybrid complexes. The main figure is the DSC melting curve, excess heat capacity change, ΔC_p, versus temperature, for the heteromorphic duplex, III_P-IV_T, detected in the optical melting experiments depicted in Figure 3. The inset plot shows the DSC melting curves for the perfect duplex, IV_T-IV_P, and the heteromorphic complexes, IV_P-III_T and III_P-IV_T. Curves are denoted by the keys on the figures. The total strand concentration is equal at ~120 μM in all three cases. Clearly, transitions detected in optical melting experiments are also apparent on the DSC melting curves and have significant thermodynamic stability. Thermodynamic parameters evaluated from these DSC melting curves are summarized in Table 2.

Table 2. Thermodynamic Parameters Evaluated from DSC Melting Curves for the Measurements for Duplex Complexes that Form between Different Pairs of Strands IV_P, IV_T, III_T, and III_P

	ΔH (kcal mol^{-1})	ΔS (cal K^{-1} mol^{-1})	ΔG (kcal mol^{-1})	t_m (°C)
$IV_P + IV_T$	201.8 ± 2.1	609 ± 6	20.1 ± 0.2	60.3 ± 0.0
$IV_P + III_T$	50.6 ± 1.7	172 ± 5	−0.6 ± 0.1	21.6 ± 0.3
$III_P + IV_T$	76.8 ± 3.0	262 ± 12	−1.1 ± 0.9	19.6 ± 0.1

These results provide a direct measure of the thermostability of the single probe IV_P in a perfect or heteromorphic duplex complex with two different targets in two separate reactions. The transition displayed by the III_P-IV_T mixture clearly shows the cross-hybridization propensity is maintained when the probe and target strands are interchanged. Again, this is analogous to the situation that might occur in a real multiplex hybridization environment.

Calculations of the Stability of the Hybrid Duplexes. Calculated values of t_m for the perfect duplexes and the hybrid mixtures, determined using the HyTher program,[40] are given in Table 1. Calculations for the perfect match duplexes were straightforward. The t_m value was determined according to the number and type of base pairs and their nearest-neighbors. For the hybrid mixtures, the calculation involved several assumptions. First, duplex structures of potential complexes were assumed, and then thermodynamic parameters for tandem mismatches were estimated as described earlier (Figure 2b). To begin these calculations, each pair of sequences to be considered were input into the primer walk option of the HyTher program and the most stable alignment and corresponding t_m of the heteromorphic state of that alignment were determined. Comparisons of the calculated and experimental results in Table 1 reveal t_m values for the perfect duplexes are reasonably predicted by HyTher (within 2.1 °C). For the hybrid mixtures, II_P-I_T and I_P-II_T, predicted t_m values for the most likely alignments are 3.6 °C and 7.3 °C higher, respectively, than those observed

experimentally. From these predictions, the likelihood that the hybrid mixtures would probably cross-hybridize with sufficient stability is evident. Using our nearest-neighbor parameters, the predicted t_m values are 23.5 °C and 22.1 °C lower than those experimentally observed for the I_T-II_P and II_T-I_P complexes, respectively. In other words, we predict a much less stable complex and Hyther predicts a more stable complex than those observed for the I_T-II_P and II_T-I_P complexes. This apparent disparity is likely due to the inclusion of a stabilizing factor for tandem mismatches used in the Hyther calculation.[40] As will be seen, when we include a stabilizing factor for tandem mismatches, a greatly improved agreement with experiment is obtained. In contrast, for the IV_P-III_T and III_P-IV_T mixtures HyTher predicts t_m values approximately 47 °C and 15 °C lower, respectively, than those observed experimentally! In contrast to the other strand mixtures, these predicted t_m values are so low compared to those of the perfect duplexes that these mixtures would be predicted not to cross-hybridize. Experimentally, these hybrid duplexes are much more stable than predicted by HyTher. Similarly, our calculations for the IV_P-III_T and III_P-IV_T complexes are approximately 38 °C and 31 °C lower, respectively, than those experimentally observed.

The aforementioned calculations indicate mixtures predicted not to cross-hybridize are experimentally observed to do so. This reveals a shortcoming of the conventional approach (represented by the HyTher calculation). As an alternative, we applied the CFD approach to analyze the sequence mixtures. This process suggests a new source of sequence dependent thermodynamic stability in heteromorphic duplex complexes. Following are steps in the analytical procedure that was employed. (1) The CFD of each pair of strands that were melted, that is, the perfect duplexes and hybrid mixtures, were constructed in two different ways. Initially, each CFD was constructed using published values of the hydrogen bonding contributions (A−T or G−C) and nearest-neighbor stacking interactions published by Benight and co-workers[29,36] used in combination with the nearest-neighbor dependent single base pair mismatch values published by SantaLucia and co-workers.[37−40] For convenience, the Benight parameters for the nearest-neighbor sequence dependent calculations were used here, since they have been shown several times to be comparable to SantaLucia's[29,30] and would not be expected to yield incomparable results at the level they are employed. For each alignment, the quantitative nearest-neighbor parameters were used to predict the respective t_m. The CFDs for the perfect match duplexes are shown in Figure 6. On the CFDs for the perfect match duplexes, a maximum occurs at the perfect alignment position that is in every case within 3 °C of the experimental t_m. This is comparable to the predictions obtained for the perfectly matched duplexes with HyTher. This also verifies that the Benight and SantaLucia nearest-neighbor sets produce comparable predictions.

For the hybrid mixtures, sequences were aligned in the various possible heteromorphic configurations, and for each of these, the t_m (and corresponding thermodynamic transition parameters) were calculated by counting the number of complementary base pairs, nearest-neighbor stacking interactions and base pair mismatches present at each alignment position. This process was continued until the state with the maximum number of base pairs and highest predicted t_m was formed, corresponding to a
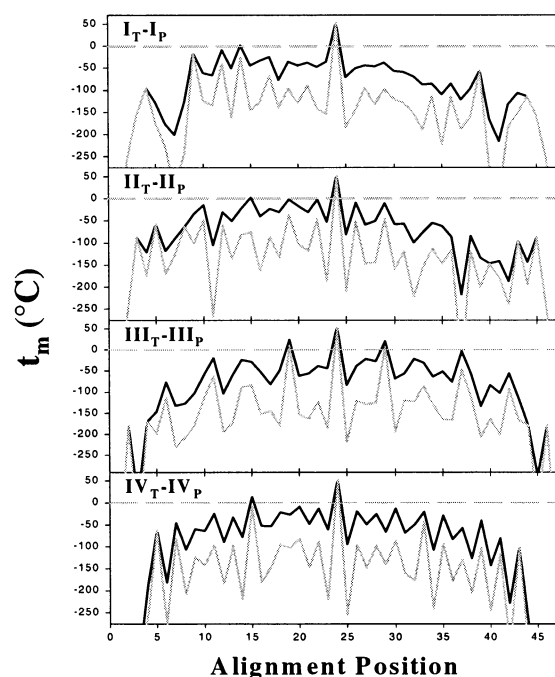

Figure 6. Conventional (gray lines) and optimized (solid lines) context functional descriptors, CFDs, for each of the perfect match probe−target complexes displayed in terms of calculated t_m values. These plots were constructed using the calculated t_m values of the various hybrid complexes encountered at each strand alignment position, as described in Figure 2. For each duplex, the calculated t_m at the alignment position of maximum overlap is within 2 °C of the experimentally determined t_m.
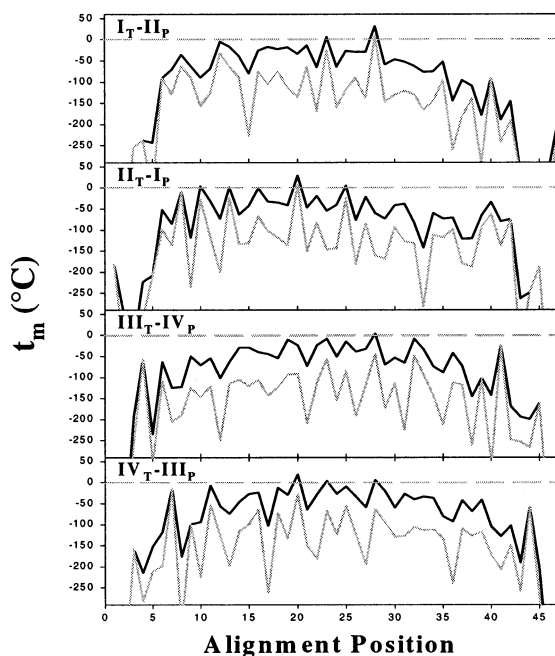

Figure 7. Conventional (gray lines) and optimized (solid lines) context functional descriptors (CFDs) calculated for the four heteromorphic duplex complexes. Note considerations regarding the stability of tandem mismatch states used to calculate the optimized CFDs (us optimized in Table 1) produce a significant difference from the CFDs calculated in the conventional sense that neglect these contributions.

maximum on the CFD, with a relative height corresponding to the calculated t_m. The CFDs for the hybrid mixtures are shown in Figure 7 (solid lines). When the standard prescription for calculating stability, that is, considering only hydrogen bonding

Table 3. Statistical Similarity between Calculated CFD Profiles for Different Pairs of Strand Mixtures

CFD profile similarity correlation coefficients, R		
duplex profile comparison	us (conventional)[a]	us (optimized)[b]
I_T-I_P vs II_T-I_P	0.548 71	0.826 72
I_T-I_P vs I_T-II_P	0.500 19	0.806 76
II_T-II_P vs II_T-I_P	0.545 42	0.837 62
II_T-II_P vs I_T-II_P	0.471 75	0.806 711
III_T-III_P vs IV_T-III_P	0.652 58	0.909 85
III_T-III_P vs III_T-IV_P	0.600 78	0.893 88
IV_T-IV_P vs IV_T-III_P	0.605 92	0.912 29
IV_T-IV_P vs III_T-IV_P	0.536 27	0.895 23

[a] Values determined by comparing CFDs constructed not considering potential stabilizing interactions in tandem mismatch states. [b] Values determined by comparing CFDs constructed considering contributions from stabilizing sequence specific interactions in tandem mismatch states.

contributions (A−T or G−C), nearest-neighbor stacking interactions, and nearest-neighbor dependent single base pair mismatches in hybrid duplex states was employed to calculate t_m, our procedure produced calculated t_m values (us, third column in Table 1) quite comparable to those predicted by HyTher.

Neither the primer walk option of the HyTher program nor our conventional approach produced calculated stabilities of the heteromorphic complexes in agreement with experimental observations. The degree of similarity between the CFD of the hybrid mixtures and the CFD of the reference (perfect matched duplex) is used as the predictor of the propensity for heteromorphic complex formation (cross-hybridization) in the hybrid mixtures. Practically, this depends on the quantitative quality and features of the CFDs constructed for the heteromorphic duplexes considered. Results of the quantitative comparisons of the various hybrid CFDs with those of corresponding perfect match duplexes are summarized in Table 3. In this comparison, the correlation coefficients were employed. When the CFDs are calculated in the conventional manner, the correlation coefficients, R,

$$R_{1,2} = \frac{1}{N_{1,2}}\sum_{P=1}^{N}(\text{CFD}_1)_P(\text{CFD}_2)_P \qquad (7)$$

range from 0.47 to 0.65. Since experimental t_m values were much higher than those predicted by the standard method using HyTher or the Benight parameters, additional stabilizing interactions, not considered in the standard method, were considered and included in calculations of stability.

The hybrid mixtures IV_P-III_T and III_P-IV_T showed the greatest departure from the conventional predictions, so, for the following example, the III_P-IV_T heteromorphic duplex is considered. For the III_P-IV_T pair of strands, the standard alignment procedure was performed and the specific alignment that produced the most hydrogen bonds between complementary base pairs was denoted. This aligned state is shown in Figure 8. Examination of this state and its sequence suggested immediately a possible source of the observed much higher than expected stability, that is, interactions of complementary bases within mismatch loops comprised of two or more adjacent base pair mismatches. To test this hypothesis, the conventional stability calculation was augmented to consider this potential source of added stability. Specifically, where adjacent mismatches occurred and bases in adjacent positions on opposite strands were complementary, a stabilizing factor equal to a
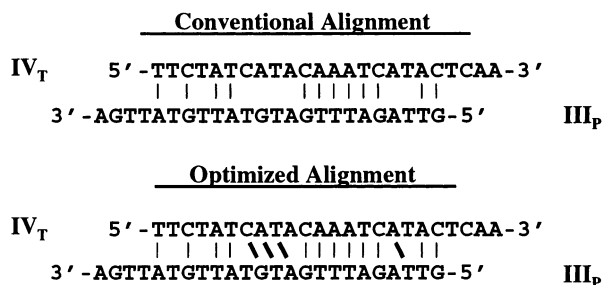
Conventional Alignment

IV$_T$ 5′-TTCTATCATACAAATCATACTCAA-3′
            | | | || |||||| ||
     3′-AGTTATGTTATGTAGTTTAGATTG-5′   III$_P$

Optimized Alignment

IV$_T$ 5′-TTCTATCATACAAATCATACTCAA-3′
            | | || \\\ |||||| \ ||
     3′-AGTTATGTTATGTAGTTTAGATTG-5′   III$_P$

Figure 8. Conventional and optimized alignments for the proposed heteromorphic complex formed between strands IV$_T$ and III$_P$. Consideration of potential stabilizing interactions between members of different mismatch base pairs in tandem mismatch loops (diagonal black bars) is the added source of stability assigned to tandem mismatch states in the optimized alignment.

fraction (0.6) of the total hydrogen bonding stability of a base pair was assigned to each occurrence. As shown in the optimized alignment Figure 8 (bottom figure), there are four such additional interactions (depicted by dark bars) that might contribute added stability to the hybrid duplex complex.

With this additional thermodynamic contribution, a new set of CFDs was constructed for the hybrid duplexes. These plots are shown in Figure 7 (solid lines) and compared directly to the CFDs constructed in the conventional way (gray lines). The calculated t_m values from these new CFDs for the hybrid mixtures are summarized in the far right column of Table 1. Also note the conventional and optimized t_m values are the same for the perfect match duplexes that do not have tandem mismatch loops. Clearly, better agreement with experimental measurements is obtained. The newly constructed CFDs of the hybrid mixtures were quantitatively compared with the CFDs of the corresponding perfect match duplexes. The resulting correlation coefficients are summarized in Table 3 under the column us (optimized). As can be seen, these values increase dramatically to a range from 0.81 to 0.91 when the CFDs are constructed, considering the suggested intramismatch loop stabilizing interactions. This means that the newly calculated CFDs of the hybrid duplexes are more similar to those of the perfect match duplexes and now, consistent with experimental observations, would be expected to have a higher propensity for cross-hybridization.

In the above analysis, the presence of additional stabilizing sequence dependent interactions within tandem mismatch loops was considered and provided calculated CFDs that were more similar to that of the perfect match duplex and, thus, would be expected to cross-hybridize as experimentally observed for the sequences examined. This suggests that the tandem mismatches might provide for considerable stabilizing interactions. Thus, our conventional estimates on the stability of tandem mismatches provide only lower limit estimates on the actual stability of tandem mismatch states. Clearly, further studies of tandem mismatches will be required to verify the existence of stabilizing interactions within tandem mismatches.

There are two obvious, very important practical implications for these findings. First, results suggest that thermodynamic stabilizing interactions might occur in internal mismatch loops comprised of more than two base pair mismatches. Second, inclusion of this conjectured new thermodynamic interaction provides better semiquantitative assessments of cross-hybridization propensity. We should add that the present method does

PM duplex
5'-TAA AAG ATA CCA TCA ATG AGG AAG CTG CAG A-3'
3'-ATT TTC TAT GGT AGT TAC TCC TTC GAC GTC T-5'

AAG/TCC (Left Side)
5'-TAA AAG ATA CCA TCA ATG AGG AAG CTG CAG A-3'
3'-ATT TCC TAT GGT AGT TAC TCC TTC GAC GTC T-5'

AAG/ T<u>C</u>C (Right Side)
5'-TAA AAG ATA CCA TCA ATG AGG AAG CTG CAG A-3'
3'-ATT TTC TAT GGT AGT TAC TCC TCC GAC GTC T-5'

Figure 9. Sequences of the 31 base pair duplex DNAs that were studied previously.[32] Positions of single base pair mismatches (bold) and flanking nearest-neighbor sequences are denoted by the boxes.

not consider effects of differing size loops in opposing strands. In addition, during the annealing process, certain pairing preferences may be kinetically favored, and these are not considered in our thermodynamic model.

Analysis of Context Effects on Single Base Pair Mismatch Stability. The CFD approach was designed to analyze heterogeneous mixtures, as discussed previously, that might occur in multiplex reactions where several or many different duplexes are present. In this case, the sequence context refers not only to the order and composition of base pairs in each perfect duplex but also to the sequences and contexts of the other strands and their relative complementarity with respect to the perfect duplexes. We surmised the CFD approach might also be applied to analyze effects of context in homogeneous mixtures of strands, defined as those where only two strands are present with sequences that are maximally complementary to each another. In this case, context refers to the explicit order and identity of all base pairs in fully duplex states. Following is an example of the application of the CFD method to analyze sequence context effects in such a homogeneous solution of duplex molecules.

Recently, we reported for a homogeneous mixture that the thermodynamic stability associated with a single base pair mismatch in a short (31 base pair) deoxyoligonucleotide duplex depends on the sequence environment, or context (beyond nearest-neighbors), surrounding the mismatch.[32] In that study, significant differences in stability were found for duplex DNAs having the same sequence with the same single base pair mismatch (A/C), flanked by the same neighboring base pairs on the 5' and 3' side, (AAG /TCC), at different positions in the duplex. The sequences of the molecules that were studied and melted in 55 mM Na+ are shown in Figure 9. At the different positions, the mismatch and its flanking nearest-neighbor base pairs reside within different sequence contexts. The CFD approach was applied to attempt to quantify subtle differences in contexts of the sequences that affect the stability of single base pair mismatches in the same nearest-neighbor sequence environment.

The CFDs for the molecules in Figure 9 were constructed by aligning two strands of a duplex (5'−3'/3'−5') and sliding one over the other, one base at a time, base by base (as shown in Figure 2a) and estimating a parameter value that characterizes the overlapping segment in each alignment. Each point on the CFD corresponds to the estimated ΔH of the state at each alignment position. On the CFD for two complementary strands that form a perfect matched duplex or a single base pair mismatch, the maximally base paired alignment position cor-
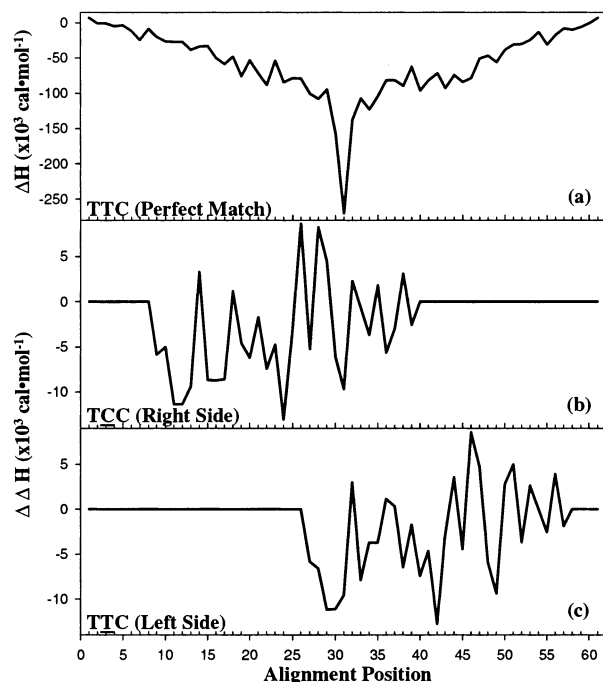

Figure 10. Context functional descriptors (CFDs) and their differences for each of the three duplexes in Figure 9. These plots are displayed in terms of the calculated transition enthalpy, $\Delta\Delta H$, versus strand alignment position. (a) The CFD for the perfect match duplex. (b) The enthalpy difference, $\Delta\Delta H$, between the CFD of the perfect match in part a and the CFD calculated for the duplex having the mismatch on the right side. (c) The enthalpy difference, $\Delta\Delta H$, between the CFD of the perfect match in part a and the CFD calculated for the duplex having the mismatch on the left side.

responds to the global minimum. The CFD constructed for the perfect match duplex sequence in Figure 9, constructed using the calculated values of the enthalpy, ΔH, at each alignment is shown in Figure 10a. Differences in the CFDs of the mismatch duplexes and the CFD of the perfect match duplex are shown in Figure 10b and c.

The global minimum at position 31 on the CFD in Figure 10a corresponds to the calculated transition enthalpy, ΔH, of the perfectly aligned and maximally base paired, perfect match duplex. The other local minima along the CFD might, in the case of heterogeneous hybridization described earlier, actually correspond to alternate heteromorphic duplex complexes. But, they need not necessarily correspond to other partially base paired duplex configurations, with corresponding ΔH values, which actually form at particular sequence alignments. Instead, in the present case of homogeneous hybridization, the CFD might be viewed as a semiquantitative representation of sequence context, not an exact thermodynamic characterization of hypothetical hybrid duplex structures.

To emphasize differences in the CFDs of the perfectly matched duplex and the mismatched duplexes, the differences in the CFDs are plotted in terms of $\Delta\Delta H$ versus alignment position in Figure 10b and c. The relative differences, as emphasized in these figures, might serve as an analytical tool for comparisons of subtle differences in contexts for the different duplexes. Clearly, the overall difference plots are similar in shape (only shifted), but significant local differences are also seen.

The CFDs for the perfect match duplex and mismatch duplexes in Figure 9 were also constructed in terms of the

Table 4. Experimental and Calculated t_m Values for the Perfect Match (PM) and Mismatch Duplexes, MM(L) and MM(R), and Statistical Analysis of the Differences in Their CFDs

duplex	t_m (°C) (exptl)[a]	t_m (°C) (HyTher)[b]	t_m (°C) (us optimized)[c]	$\Delta t_m{}^d$ (°C)	$\Delta t_m{}^e$ (°C)	$\Delta t_m{}^f$ (°C)	$\Sigma(\delta^2)^g$
PM	62.0	62.4	62.5	0	0	0	0.0
MM(L)	60.9	58.3	60.9	−1.1	−4.1	−1.6	1140.5
MM(R)	57.4	58.3	60.7	−4.6	−4.1	−1.8	1356.6

[a] t_m values evaluated from experimental melting curves reported in ref 32. [b] t_m values calculated from HyTher program.[40] [c] t_m values calculated using parameters reported in ref 36 with the stabilizing contribution of tandem mismatches. [d] $\Delta t_m = t_m(MM) - t_m(PM)$ for t_m(exptl). [e] $\Delta t_m = t_m(MM) - t_m(PM)$ for t_m(HyTher). [f] $\Delta t_m = t_m(MM) - t_m(PM)$ for t_m(us optimized). [g] Sum of the squares of the differences between our optimized CFDs from eq 8.

calculated t_m of each aligned state (not shown). Quantitative differences between these CFDs were assessed by calculating the sum of the squares of the differences, $\Sigma\delta^2$,

$$\sum_{P=1}^{N}\delta^2 = \sum_{P=1}^{N}[(CFD_2)_P - (CFD_1)_P]^2 \qquad (8)$$

between the respective CFDs. As summarized in Table 4, the value of $\Sigma\delta^2$ is significantly smaller for the mismatch duplex, AAG/TCC (L), which also happens to be relatively more stable with a stability closer to that of the perfect match, PM, duplex. For these data, the greater the difference in t_m values between a mismatch and perfect match duplex, the larger the values of $\Sigma\delta^2$ between their respective CFDs. In addition, Δt_m of the mismatch duplexes scales linearly with $\Sigma\delta^2$. This further demonstrates the quantitative significance of the CFD approach and how this method might provide a novel way to quantitatively characterize DNA sequence context.