

Hybridization properties of long nucleic acid probes for detection of variable target sequences, and development of a hybridization prediction algorithm

Christina Öhrmalm¹, Magnus Jobs^{1,2}, Ronnie Eriksson¹, Sultan Golbob¹, Amal Elfaitouri¹, Farid Benachenhou¹, Maria Strømme³ and Jonas Blomberg^{1,*}

¹Clinical Virology, Department of Medical Sciences, Uppsala University and Academic Hospital, 751 85 Uppsala, ²School of Health and Social Sciences, Högskolan Dalarna, Falun and ³Nanotechnology and Functional Materials, The Ångström Laboratory, Uppsala University, Box 534, 75121 Uppsala, Sweden

Received February 13, 2010; Revised August 10, 2010; Accepted August 15, 2010

ABSTRACT

One of the main problems in nucleic acid-based techniques for detection of infectious agents, such as influenza viruses, is that of nucleic acid sequence variation. DNA probes, 70-nt long, some including the nucleotide analog deoxyribose-Inosine (dInosine), were analyzed for hybridization tolerance to different amounts and distributions of mismatching bases, e.g. synonymous mutations, in target DNA. Microsphere-linked 70-mer probes were hybridized in 3M TMAC buffer to biotinylated single-stranded (ss) DNA for subsequent analysis in a Luminex[®] system. When mismatches interrupted contiguous matching stretches of 6 nt or longer, it had a strong impact on hybridization. Contiguous matching stretches are more important than the same number of matching nucleotides separated by mismatches into several regions. dInosine, but not 5-nitroindole, substitutions at mismatching positions stabilized hybridization remarkably well, comparable to N (4-fold) wobbles in the same positions. In contrast to shorter probes, 70-nt probes with judiciously placed dInosine substitutions and/or wobble positions were remarkably mismatch tolerant, with preserved specificity. An algorithm, NucZip, was constructed to model the nucleation and zipping phases of hybridization, integrating both local and distant binding contributions. It predicted hybridization more exactly than previous algorithms, and has the potential to guide the design of variation-tolerant yet specific probes.

INTRODUCTION

Microbial genomes can be highly variable because of high mutation rates. Because of this extreme variability, it is often difficult to identify regions within a specific virus genome that are sufficiently evolutionarily conserved to serve as targets for specific detection primers and probes. RNA viruses are especially variable. The influenza virus, a negative sense, single-stranded RNA (ssRNA) virus with a highly variable RNA genome, for example, has been known to cause the diagnostic problem that is at the basis of this article, because of a high rate of mutation and genetic drift. In such situations, optimal detection primers and probes would be broadly targeted yet specific, and remain functional even if the genome sequence changed because of genetic drift.

Diagnostic nucleic acid hybridization probes are constructed from the most conserved portions of genes from viruses commonly causing infection. Long probes have a large inherent tolerance to microbial variation. The introduction into the introduction into the probe design of a base that can hybridize with all four normal bases (a universal base), or of multiple nucleotides (degenerations; wobbles) in a single position, can induce tolerance to natural viral variation (mismatch).

The naturally occurring (1–4) nucleotide (nt) deoxyribose-Inosine (dInosine) is one of many more or less generally hybridizing nt (5,6) known as universal bases. All four normally occurring DNA bases can hybridize to dInosine. The general trend in decreasing hybridization stability is I:C > I:A > I:T ≈ I:G > I:I when using 1 M NaCl, 10 mM sodium cacodylate and 0.5 mM EDTA pH 7 (7,8). However, dInosine is readily available and can be recognized as a G by polymerases (5,9,10). Alternative universal bases, e.g. 5-nitroindole, also exist.

*To whom correspondence should be addressed. Tel: +46 18 611 5593; Fax: +46 18 551 012; Email: jonas.blomberg@medsci.uu.se

3 M tetramethylammonium chloride buffer (TMAC) is a hybridization buffer that selectively raises the stability of A:T base pairs to approximately that of G:C base pairs (11–14). It was used in these studies to reduce the effect of sequence composition when comparing different probes of the same length. The term nucleation site is used in this article to indicate a stretch of contiguous perfectly matching nt, capable of initiating hybridization (15).

The aim of these studies was to improve understanding of the design of probes to be used in a TMAC buffer system, by investigating variability and the inclusion of dInosines, other universal bases and wobbles. Specifically, we examined (i) variation (i.e. mismatch) tolerance, (ii) sensitivity to different mismatch distributions, (iii) utilization of dInosine as an nt analog, and (iv) specificity; we also present a new algorithm for prediction of hybridization results. In additional experiments, the question of the use of degeneracy versus a universal base was addressed. Furthermore, we investigated the use of the derived design criteria for detection of rotavirus RNA in clinical samples.

MATERIAL AND METHODS

A 70-mer nucleic acid hybridization probe, named the InflA probe, was constructed from the most conserved portion of the matrix gene in segment 7 of the Influenza A H3N2 virus. Properties important for the design of variation (mismatch)-tolerant yet specific probes were investigated by studying the interaction between a set of virus-derived probes and complementary targets with different degrees and distributions of mismatch.

The 70-mer DNA probes were coupled to color-coded microspheres, hybridized with biotinylated target nucleic acids, incubated with streptavidin–phycoerythrin and analyzed in a Luminex® 200™ system. The hybridization reaction was performed at a standard non-saturating concentration (0.2 nM target) in 3 M TMAC buffer.

In further experiments, in an attempt to make a probe with an extended mismatch tolerance, a series of dInosine-containing probes were synthesized and analyzed for hybridization in the 3 M TMAC buffer system. A limited number of experiments comparing dInosine with an alternative universal base (5-nitroindole) or with wobbles (degenerations) were also conducted.

Many viruses harbor synonymous mutations (sm), which means that the third base in a codon can wobble without changing the amino acid in the protein. Targets with regions where every third base was mutated, to resemble the common phenomenon of sm (i.e. the nt sequence is varied without affecting the coding for a specific amino-acid code), were of special interest, since this is often the cause of the variation in coding viral sequences.

Rotavirus is a dsRNA virus that causes gastroenteritis. Clinical fecal samples, previously confirmed to contain rotavirus, were used to test the design strategies of this report. An asymmetric PCR, with biotinylated rev-primer in excess, was set up for the VP6 segment. The primers

were designed using knowledge gained from this report, i.e. both degeneration and dInosine were used to create moderately degenerated probes with uninterrupted matching stretches as long as possible, in the most conserved regions. The biotinylated reverse (rev-) primer was made 77-nt long to be as tolerant to variation as possible, while the shorter forward (fw-) primer (23 nt) contained four locked nucleic acids (LNA) to increase the hybridization strength (represented by melting temperature, T_m) to match that of the rev-primer.

Probe design

Single stranded 70-mer oligonucleotides with a C12 aminolink at the 5' end, with or without dInosines/5'-nitroindole/N wobbles, were obtained from Biomers.net (Ulm, Germany). The design of the probes was based on the programs BLASTn (16), ClustalX (17) and ConSort® (J.Blomberg *et al.*, unpublished results). Briefly, matching viral sequences were retrieved from the GenBank database at NCBI, NIH, using the viral sequence of interest as a query, and alignments were performed by a BLASTn search. BLASTn and ClustalX alignments were analyzed in ConSort® to define the most suitable probe sequence. ConSort® provides the frequency of variation and the variation of nt composition in each base position, the number of aligned sequences, and a majority consensus sequence. The proposed probe sequence was further analyzed for its predicted T_m and probable homodimer and hairpin interactions using Mfold at the IDT OligoAnalyser site (<http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/>) [<http://mfold.bioinfo.rpi.edu/> (18)] and Visual OMP™ 7.0 (DNA Software). Visual OMP™ 7.0, which uses the Nearest Neighbor (NN) algorithm [DNA Software (19)], was used to estimate the change in Gibb's free energy associated with hybridization of the two strands (ΔG) for the interaction between probes and targets at 45 and 55°C in 3 M TMAC. IDT OligoAnalyser was used to estimate ΔG for the interaction between probes and targets using 50-mM Na⁺ and 2-mM Mg²⁺. Each combination was tested with the heterodimer-formation function (<http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/>).

To create the 70-nt InflA probes, nt 725–794 of the matrix protein 2 gene segment 7 of Influenza A H3N2, accession nr CY023083, was used as query in BLASTn (GenBank database at NCBI, NIH). The Norovirus probe sequence comprises nt 646–715 of the capsid gene of the Norwalk-like virus, accession nr AY274264. The probes for detection of rotavirus were made using sequence EU372725 of the Human rotavirus A strain CMH171/01 inner capsid protein (VP6) gene as query. The probe region chosen, nt 112–178, was analyzed using the haplotype function of ConSort®. Two differently degenerated probes containing dInosines were designed from each of the two major haplotypes, which potentially covered all rotavirus group A variations recorded in GenBank.

Synthetic targets, complementary to the consensus sequence of the InflA, Norovirus and Rotavirus probes,

with various numbers of mismatches, were purchased as 70-mer oligonucleotides with biotin attached to a 2-aminoethoxy-ethoxyethanol linker at the 5' end (Biomers.net, Ulm, Germany).

Probe coupling to microspheres

Specific synthetic 5' amine-C12 modified 70-mer probes for influenza A were designed and solid-phase coupled to xMAP carboxylated color-coded microspheres (Luminex Corp., Austin TX, USA), according to the protocol of the Luminex corporation (Austin TX, USA). Briefly, 2.5×10^6 stock microspheres were collected by centrifugation and resuspended in 25 μ l of 0.1 M MES, pH 4.5 (2*N*-morpholino-ethanesulfonic acid, Sigma). Subsequently, 0.2 nmol of the probe and freshly made 2 μ l 10 mg/ml EDC [1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (water-soluble carbodiimide; Pierce; sold by Nordic Biolabs AB, Sweden)] in H₂O were added to the microspheres and the suspension was incubated in the dark for 30 min at room temperature. Care was taken to store the EDC in a dry condition, in aliquots. After addition of another 2 μ l (10 mg/ml) EDC in H₂O and repeated incubation, the microspheres were washed with 0.5 ml of 0.02% Tween-20. The coupled microspheres were pelleted by centrifugation at $8000 \times g$ for 2 min and resuspended in 0.5 ml of 0.1% SDS. After a second spin at $8000 \times g$ for 2 min, the final pellet was resuspended in 50 μ l of TE, pH 8.0.

Hybridization of biotinylated target DNA to probe linked microspheres

An amount of 5 μ l of 2.0 nM synthetic biotin-labeled target was mixed with hybridization buffer consisting of 33 μ l 3 M TMAC buffer (3 M tetramethylammonium chloride, 0.1% Sarkosyl, 50 mM Tris-HCl, pH 8.0, 4 mM EDTA, pH 8.0; Sigma) and 12 μ l 1 \times TE-buffer pH 8.0 and 0.05 μ l (~2500 microspheres) for each probe-coupled Luminex microsphere. The mixture was heated at 95°C for 2 min to denature the DNA targets and probes, followed by hybridization at 45 or 55°C for 30 min while shaking on the ThermoStar (BMG LabTech; Offenburg, Germany) microplate incubator. An amount of 2 μ l (0.05 mg/ml) of streptavidin-R-phycoerythrin (QIAGEN, Hilden, Germany) was added to the mixture, which was further incubated at 45 or 55°C for 15 min before analysis for internal microsphere and R-phycoerythrin reporter fluorescence on the Luminex®200™ system (Luminex corporation, Austin, TX). The amount of biotinylated target that hybridizes to the microsphere-bound probes is directly proportional to the Median Fluorescence Intensity (MFI) reported by the instrument (in all experiments, fluorescence was measured from a minimum number of each type of microsphere: a set of 100 beads). The term total MFI describes the hybridization signal from a perfectly matching probe-target duplex, while the percentage (%) of the total MFI describes the ratio between the hybridization signal of a mismatching probe-target duplex and the total MFI of that particular probe. Titration experiments established that ~60% of the maximum hybridization capacity of

the microsphere-bound probes was reached using the conditions under which total MFI was measured (i.e. microsphere-bound probes were not saturated). An MFI of 100 was used as the lower limit of detection (LLOD).

Clinical samples containing Rotavirus

Fecal samples were obtained from children with gastroenteritis, from the Children's Hospital ward at Uppsala Academic Hospital. Samples were handled anonymously according to the rules of the ethical committee at the Academic Hospital. The study used samples that were positive in a rotavirus antigen detection test; 100 μ l of the sample was diluted in 900 μ l 1 \times TE buffer. After centrifugation, 400 μ l was added to a lysis buffer and total nucleic acid was extracted as described by the manufacturer (easyMag®, bioMérieux). The samples were eluted in 110 μ l buffer and stored at -70°C.

Asymmetric RT-PCR and sequencing of Rotavirus dsRNA in clinical samples

A reverse transcriptase (RT)-PCR was set up to amplify and biotin-label the nucleic acid of human rotavirus A from these clinical samples.

The 545 sequences obtained from the rotavirus query in BLASTn were analyzed in the ConSort® program to construct fw- and rev-primers. The fw-primer (nt 1-23), 5'-G GCTTTW+AAA+CGAA+GTC+TTCR-3' (+A, +C, +G, +T are LNA residues) and the biotinylated rev-primer (nt 502-426), 5'-TATGGAAATATATTAGG TTTATGAAAAACAAATCCIGTACGTTGTCTTCT ITTITGIARRTTCCAITTITCIATRTA-3', resulted in a PCR product of 502 nt. After nucleic acid extraction in the easyMag®, the extracts were heated at 97°C for 5 min, followed by snap cooling on ice for 2 min to obtain ssRNA from the rotaviral dsRNA. The PCR reaction contained 5 μ l of nucleic acid extract, 25 μ l 2 \times RT-PCR iScript buffer, 200-nM Fw-primer, 600-nM biotinylated rev-primer, nuclease-free water and 1 μ l iScript RT enzyme (total volume 50 μ l). The samples were run at 50°C for 30 min, 94°C for 10 min, 50 cycles at 94°C for 30 s, 55°C for 30 s and 72°C for 30 s, ending with 72°C for 7 min. An amount of 5 μ l of the PCR product was used in the hybridization experiment with the microsphere-coupled Rotavirus probes (sequences shown in Supplementary Table S4A and B). The PCR products were sequenced with a 3130 Genetic Analyzer (Applied Biosystem) by utilizing the same fw- and rev-primers described above.

Development of a new algorithm

The discovery of the importance of long uninterrupted perfect matches and the long-range effects of mismatch (see 'Results' section), which are not embodied in current NN hybridization theories (19), led us to formulate a simple new descriptive theory. The new algorithm includes aspects of NN theory (8,20-22) but extends this to longer hybridizing segments and includes the effects of dInosine in the long oligonucleotides. Visual Omp™ 7.0 predicted that some 70-mer combinations would not

hybridize, yet they did hybridize. In order to better predict hybridization, we investigated predictive strategies that took into account the matching nt, its neighbors, the length of the matching region, and the cooperativity of neighboring matching regions. We finally settled for an algorithm which attempts to model the nucleation and zipping stages during the hybridization process. The new model was termed NucZip (Figure 5B). The results obtained using this model was then correlated with those obtained using Visual OMP™ 7.0 and both were compared with experimental data.

How NucZip works

NucZip simulates the hybridization process, starting with potential nucleation sites and then proceeding to 'zip' in both directions. The NucZip algorithm (written by JB in Visual FoxPro) was implemented as a module (procedure) in the ConSort® sequence analysis program. The procedure is relatively simplistic, and does not contain the thermodynamic and secondary structure analyses provided by more sophisticated programs such as Visual OMP™ 7.0.

The algorithm starts by searching for perfectly matching hexa-, hepta-, octa- and nonamers, as potential nucleation sites (the 'Nuc' part). Every matching oligomer is given the number of matching nt as its NucScore. If the oligomer contains dInosine, the number of dInosines is subtracted from the score but the segment is still counted as uninterrupted, regardless of the dInosines. The NucScore is then used to select the two highest scoring potential nucleation sites which will undergo zipping, up and downstream.

The score for the 'Zip' portion of the model is obtained from the number of consecutive matching trimers, tetramers etc, up to pentadecamers, each counted with equal weight, within a contiguous matching segment. Thus, $\text{ZipScore}_{\text{modex}} = \sum_{k=1}^{k_{\text{max}}} (\sum_{n=3}^{15} S_n)$, where k_{max} is the number of uninterrupted matching segments, including the chosen nucleation site, and S_n is the number of successive segments of length n (varying from 3–15), i.e. the number of full length trimers, full length tetramers etc. up to full length pentadecamers, which fit into the matching segment. The same Zip scoring system was performed in two modes, counting dInosines either as matching (mode 1) or as mismatching (mode 2). In the second mode, dInosines will shorten the length of matching segments, decreasing the score. The final ZipScore was calculated as a weighted mean of $\text{ZipScore}_{\text{mode1}}$ and $\text{ZipScore}_{\text{mode2}}$, where the weighting factor was based on the empirical data presented in the current report (Figures 2–4). Since dInosine hybridizes more strongly to C than to the other nts, the algorithm adds a contribution based on the number of dI:C pairs weighted by an InoCfactor. The upstream and downstream ZipScores were obtained as: $\text{ZipScore}_{\text{downstream}} = \text{ZipScore}_{\text{mode1}} - \text{dInosinefactor} (\text{ZipScore}_{\text{mode1}} - \text{ZipScore}_{\text{mode2}}) + (\text{dInoCnr} * \text{dInoCfactor})$ and $\text{ZipScore}_{\text{upstream}} = \text{ZipScore}_{\text{mode1}} - \text{dInosinefactor} (\text{ZipScore}_{\text{mode1}} - \text{ZipScore}_{\text{mode2}}) + (\text{dInoCnr} * \text{dInoCfactor})$.

The ZipScores from up- and downstream zipping were then added; the final NucZipScore = ZipScore downstream + ZipScore upstream.

In this way, the contribution from longer matching segments was factored in with the contribution from the nearest neighbors (approximated by the trimer part of the algorithm). Figure 8 summarizes the principle of this computational work.

The probability of a match in a degenerated nt position is approximately predicted from the ConSort© analyses of target sequences. The behavior of the relatively few oligonucleotides with degeneration that were tested is approximately in line with NucZip reasoning, which places a premium on long uninterrupted matching stretches. However, a larger number of degenerated probes need to be analyzed before the contribution of probabilities of contiguous stretches extending through degenerated positions can be estimated and included into NucZip. The programming code is included in Supplementary Data.

RESULTS

The distribution of mismatches has a profound effect on hybridization

ConSort© was used to demonstrate the variation in the nt sequence of segment 7 from 7333 genomes of Influenza A (GeneBank database at NCBI, NIH). All H and N influenza A types, HxNy, are represented in the alignment (Figure 1). In comparison with the InflA probe designed to match a 70-nt region of the Influenza A H3N2 virus, the H5N1 virus differed in 5-nt positions, and the H1N1 virus in three other nt positions (Table 1). Thus, if a detection probe could tolerate mismatches in nine positions, including the variant nt positions of the H1N1 and the H5N1 viruses, it would fully cover 67 different H and N combinations of Influenza A, i.e. nearly all recorded variants in the chosen region, as demonstrated in the BLASTn search.

The InflA probe was tested against 70 nt target molecules with 3, 5, 7, 9, 11, 12, 13, 14, 15, 16 and 21 point mutations (pm) and 21 grouped mutations (gm) (Figure 2A–D; nt sequences of probes and targets can be found in Supplementary Tables S1A and B). The positions of the pm were based on the variations found by comparing H3N2, H5N1 and H1N1 viruses. Two targets had the same number of mutations, but different distributions: the 21 pm target had 21 evenly distributed pm, and the 21-gm target had seven groups of three mutations interspersed by 5 to 7 conserved nt. The InflA probe, coupled to Luminex microspheres, was allowed to hybridize with one of the biotinylated ssDNA targets in 3 M TMAC at two different temperatures: 45°C (Figure 2A) and 55°C (Figure 2B). The hybridization, given as MFI, was analyzed in the Luminex flow meter.

Introducing an increasing number of evenly distributed pm in the target had a negative effect on hybridization, as reflected in decreasing MFI; see Figure 2A (45°C) and B (55°C). A few mismatches (3 pm and 5 pm) in the 70-mer target did not affect hybridization with the probe to any great extent. A target with 9 pm (containing one 15-nt, one

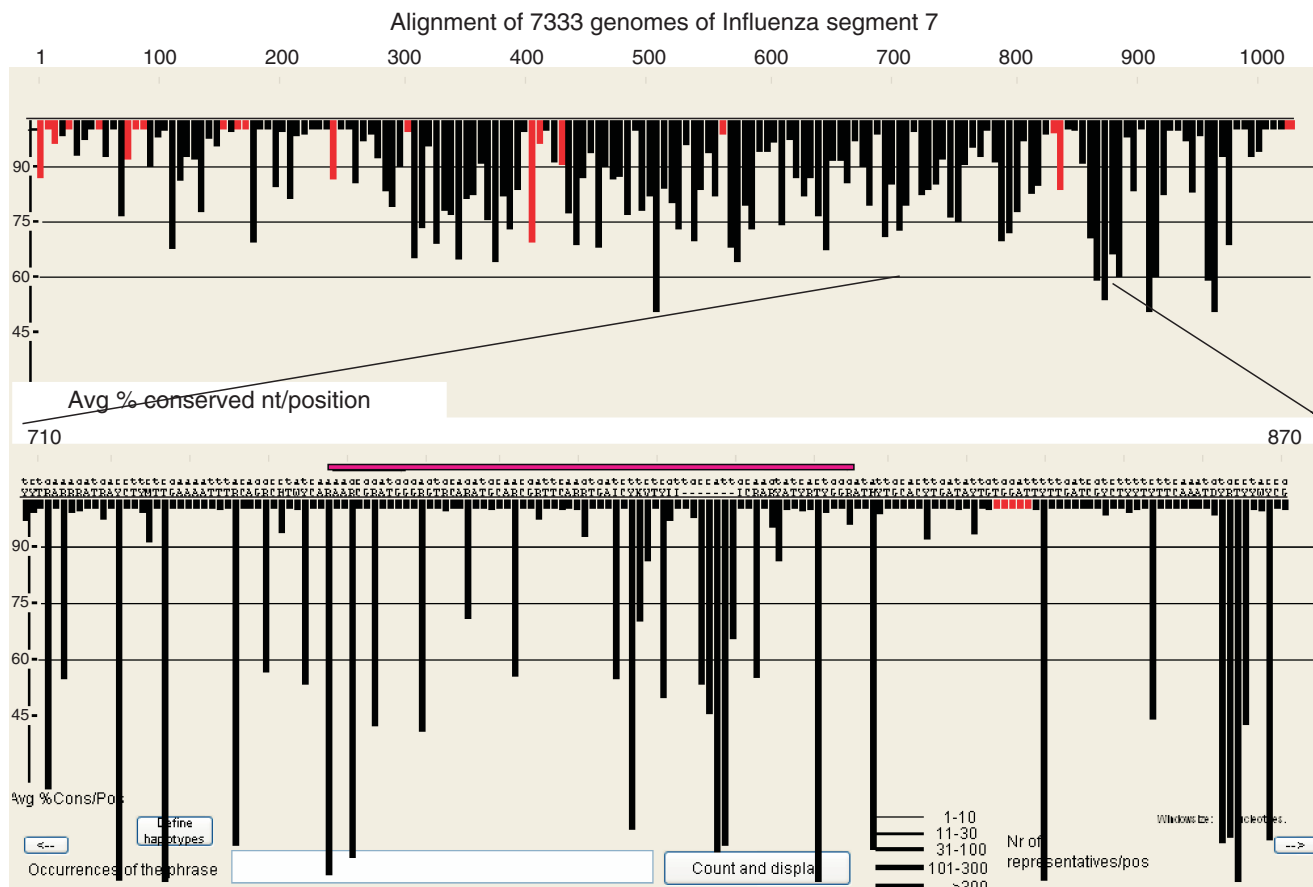


Figure 1. Alignment of 7333 genomes of Influenza A segment 7 in ConSort©; most subtypes represented. In the search for a suitable region for a 70-mer probe, the variations were mapped by BLASTn (16) and ConSort© (Blomberg,J., unpublished data). Red bars represent conserved nt, black bars represent variations from base composition and the y-axis provides the average percentage conserved nt at each position. The variation pattern is not a perfect example of a synonymous mutation pattern, where variation tends to be highest at the third base in each codon.

Table 1. The variation in a conserved region of segment 7 of the matrix gene, between subtypes H3N2, H5N1 and H1N1 of the Influenza A virus

Subtype	pm	Sequence 5'–3' (70 nt)
H3N2	0	AAACGAATGGGGGTGCAGATGCAACGATTCAAGTGACCCGCTTGTGTTGCCGCGAGTATCATTGGGATC
H5N1	5	AAACGAATGGGAGTGCAGATGCA ^g CGATTCAAGTGATCCTCTTGTGTTGCCGCAAGTATCATTGGGATC
H1N1	8	AAACG ^g ATGGGAGTGCAGATGCA ^g CGATTCAAGTGATCCTCT ^g CGTTATTGCCGCAAGTATCATTGGGATC
HxNy*	9	AAACG ^g ATGGGAGTGC ^{aa} AATGCA ^g CGATTCAAGTGATCCTCT ^g CGTTATTGCCGCAAGTATCATTGGGATC

The grey nucleotides represent the differences between the sequences. The H5N1 virus subtype differed in 5 nt and the H1N1 virus in three additional nt positions from the H3N2 virus. The sequence having 9 pm is present in 64 different subtypes (HxNy*). The results of analysis of 7333 genomes of Influenza A using BLASTn (16) and ConSort© (Blomberg,J., unpublished) are summarized. The H3N2 sequence was used as the InflA probe sequence in this article.

12-nt, one 7-nt and four 5-nt matching regions) still hybridized at 68% (MFI 4496, 45°C, Figure 2A) and 71% (MFI 4001, 55°C, Figure 2B) of the total MFI (i.e. of the MFI of the perfectly matching InflA target/InflA probe: MFI 6604, 45°C, Figure 2A; and MFI 5657, 55°C, Figure 2B) while the 15 pm target (containing one 6 nt, four 5 nt, and two 4 nt matching regions) hybridized at 13% (MFI 854, 45°C) and 1% (MFI 54, 55°C) of the total MFI. The InflA probe failed to hybridize with the 16-pm target (containing one 6-nt, three 5-nt, and two 4-nt matching regions) at either temperature, providing MFI

values that were 6% (MFI 403, 45°C) and 0.5% (MFI 29, 55°C) of the total MFI.

The longest perfectly matching sequence between the mismatches in the InflA probe/21-pm target combination was 3 nt in length. No hybridization was detected at either temperature. In contrast, for the InflA probe/21-gm target combination, where the distribution of the 21 mutations created seven stretches of 5–7 perfectly matching nt between the mismatches, hybridization with the InflA probe was restored (Figure 2A–D). The MFI signal increased to 20 and 30% (Figure 2A and C, 45°C)

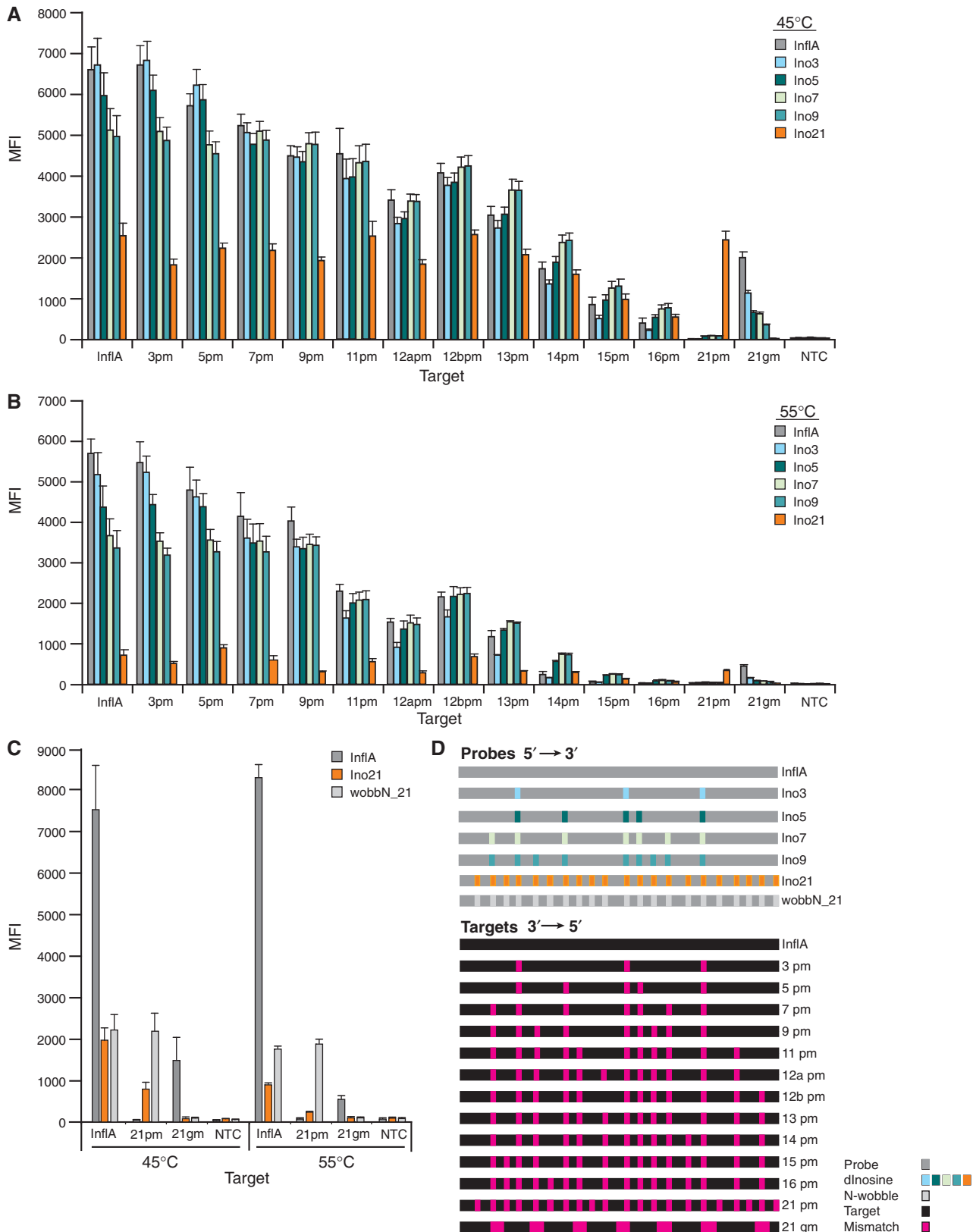


Figure 2. Hybridization of the InflA probe, with and without an increasing number of dInosines, with targets containing increasing numbers of mismatching nucleotides. The sequence of the InflA probe was designed from a conserved region of the matrix 2 gene in segment 7 of subtype H2N3 of the Influenza A virus (Figure 1, and Table 1 sequence of H3N2). The number of dInosines included in the probes is indicated in the names Ino3–Ino21. Similarly, the wobbN_21 probe contains 21 N wobbles in the same position as the dInosines in the Ino21 probe. The complementary target of the InflA probe is named the InflA target. These targets have 0–21 pm, as indicated by the names; the 21-gm target has seven groups of three mismatches (Figure 2D). All target molecules were biotinylated at their 5' end. Each sample contained 0.2 nM of one of the synthetic biotinylated

Continued

and 7 and 8% (Figure 2B and C, 55°C) of the total MFI, when the InflA probe hybridized with the 21-gm target instead of the 21-pm target.

In conclusion, a target with one to nine evenly distributed mismatches, preserving multiple contiguous matching stretches of at least 5nt, has little reducing effect on hybridization with a 70-mer probe, while a target with >14 evenly distributed pm hybridizes inefficiently or not at all. Hybridization can be improved by utilizing a less stringent (lower) hybridization temperature. Furthermore, grouped mismatches tend to strengthen the hybridization compared to evenly distributed mismatches. Consequently, for hybridization between 70-mer strands with 10–20 mismatches, the distribution of the mismatching nt affects the hybridization more than the number of mismatches, indicating that the length and number of perfectly matching stretches are of greatest importance.

Introducing dInosines into the probes at sites of variation restores hybridization with highly variable targets

A panel of five 70-mer probes, Ino3–21, containing 3, 5, 7, 9 or 21 dInosines, was designed based on the InflA probe sequence. The dInosines were placed to match the positions of the pm in the above pm targets (Figure 2D).

Introduction of 3–9 dInosines in the probe resulted in only a small reduction in MFI when binding to the InflA target; i.e. the MFI signal decreased in the order InflA probe \approx Ino3 > Ino5 > Ino7 > Ino9. The Ino9 probe hybridized with the InflA target by as much as 76% (MFI 4970, 45°C) and 59% (MFI 3340, 55°C) of the total MFI, comparable with the hybridization of the InflA probe to the 9-pm target. In fact, all the Ino3–9 probes hybridized as efficiently with targets that had up to the same number of pm as the number of dInosines, including two mismatches not covered by dInosines, as they did with the InflA target. When the Ino probes hybridized with targets containing more than 12 pm, the probes with many dInosines worked better than the InflA probe; e.g. the Ino7 and Ino9 probes resulted in a signal 1.2–1.9 times higher than that for the InflA probe for targets with 13–16 pm at 45°C.

Interestingly, even the Ino21 probe hybridized quite strongly with the InflA target, at 26 and 39% (Figure 2A and C, 45°C) and 11 and 13% (Figure 2B and C, 55°C) of the total MFI. Importantly, the Ino21 probe was able to restore hybridization with the 21 pm target (11 and 37% of total MFI) to which the InflA probe had totally failed to bind (Figure 2A and C, 45°C). The Ino21 probe hybridized with all the matching 3–21 pm targets with almost the same efficiency (mean 29.5% of the total MFI, 1946 MFI, at 45°C, Figure 2A;

and 8% of the total MFI, 435 MFI, at 55°C, Figure 2B), which is in the same range as the hybridization of the InflA probe with the 21-gm target (30.4% of the total MFI, 2005 MFI, 45°C, Figure 2A; and 7.8% of the total MFI, 444 MFI, 55°C, Figure 2B). The Ino21 probe and 21 gm target combination had 13 mismatches outside the dInosine positions and failed to hybridize (Figure 2A, B and C). In conclusion, the presence of dInosine in the probe decreased hybridization with a perfectly matching target, but a dInosine-containing probe bound more strongly than a dInosine-free probe when the target had many mismatches juxtaposed to the dInosine residues.

A minimum length of perfectly matching nt sequences is required for hybridization

The importance of uninterrupted matching regions was analyzed by making targets with different lengths of perfectly matching sequences at the 5' end or at both the 5' and 3' ends, in combination with a long region of 26, 33 or 74% randomly distributed mutations (Figure 3A, B and D, nt sequences in Supplementary Table 2A and B).

As expected, hybridization of the InflA probe to targets with 26% random mutations in a region between two flanking regions of 5 (26%5F), 7 (26%7F), 9 (26%9F) or 15 (26%15F) perfectly matching nt showed that shorter perfectly matching flanking regions reduced the MFI. All the 26% targets hybridized with the InflA probe at 45°C but, at 55°C, the 26%5F target failed to hybridize (5.6% of the total MFI, 271 MFI). Like the 33%12F and the 16-pm target, the 26%5F target contained 16 mutations (Figures 2A, B, 3A and B). The 33%12F target (45% of the total MFI, Figure 3A), with its two regions of 12 uninterrupted nt, hybridized much more strongly than the 26%5F (27% of the total MFI) or 16-pm (6.1% of the total MFI) targets, which both contained several shorter matching regions, of 6, 5, 4 and 3 nt (Table 2). The same effect, i.e. that a few long perfectly matching regions result in better hybridization than several shorter regions, was also seen with the three targets that had 14 mismatches: the 26%9F, the 33%15F and the 14-pm targets at both temperatures (Table 2). The 74%9F target, with only 12 matching nt dispersed in the central region, did not hybridize to the InflA probe (1.4% of the total MFI, 86 MFI, 45°C; and 0.6% of the total MFI, 30 MFI, 55°C), while the 26%9F target, with 38 matching nt (three regions of 5 nt and one region of 6 nt), did hybridize (44% of the total MFI, 2583 MFI, 45°C; and 24% of the total MFI, 1125 MFI, 55°C). This shows that the two flanking regions of nine matching nt did not cause hybridization alone. Furthermore, the 74%12F, with 11 matching nt dispersed in the central

Figure 2. Continued

single-stranded targets and a mixture containing 2500 microspheres of each of the unique coupled probes tested. In (A) and (B), the mixture contained the InflA probe and the Ino3, 5, 7, 9 and 21 probes; in (C), the mixture contained the InflA probe, the Ino21 probe and the wobbn_21 probe. Hybridization was performed at 45°C (A and C) and 55°C (B and C) in a buffer containing 3 M TMAC for 30 min, followed by 15-min incubation with Streptavidin-R-phycoerythrin. A total of 100 microspheres per type of probe were analyzed in the Luminex 200; the extent of hybridization is indicated as median fluorescent intensity (MFI) on the y-axis. Each experiment contained triplet samples and was repeated two or three times. The sequences of probes and targets are provided in Supplementary Table 1A and B. (D) The figure shows the distribution of dInosines (orange/light blue/green boxes) and N wobbles (light grey boxes) in the probes (grey bar), and the mismatching bases (magenta boxes) in the target molecules (black bar), used in Figure 2A–C.

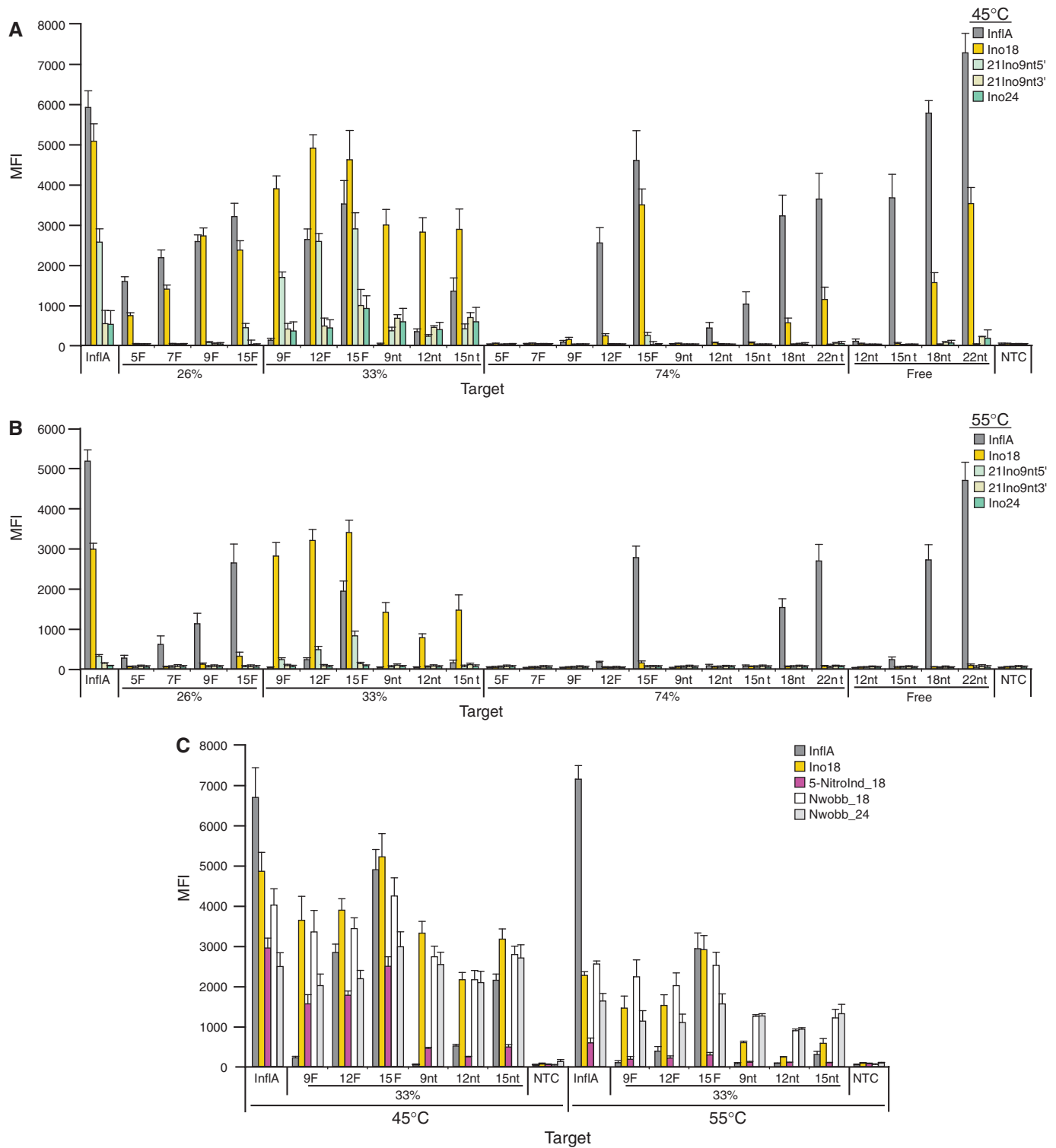


Figure 3. The length of nucleation site necessary to induce hybridization and the effect of dInosine-containing probes on sm. MFI from a microsphere mix containing 2500 microspheres of each of the probe type indicated, after hybridization with 0.2 nM of synthetic biotinylated target. Each experiment contained triplet samples and was repeated two or three times. In (A) and (B), the InfIA, Ino18, 21Ino9 nt5', 21Ino9 nt3' and Ino24 probes were hybridized at 45 or 55°C with one specific target at a time; the targets had different percentages of mismatch (26, 33, or 74%) and one or two perfectly matching regions of different length (5–22 nt) compared with the InfIA probe. The targets 12 nt_{free} to 22 nt_{free} consisted of 12–22-nt oligonucleotides. In (C), the InfIA, Ino18, 5-nitroindo₁₈, Nwobb₁₈ and Nwobb₂₄ probes were hybridized at 45 or 55°C with one specific target containing 33% mismatch and one or two perfectly matching regions of different length (9–15 nt) compared with the InfIA probe. The sequences of probes and targets are provided in Supplementary Table S2A and B. (D) The figure shows the distribution of dInosines (yellow green boxes), 5-Nitroindole (purple boxes), and wobbles of N (light grey boxes) in the probes (grey bar) and the mismatching bases (magenta boxes) in the target molecules (black bar), used in Figure 3A–C.

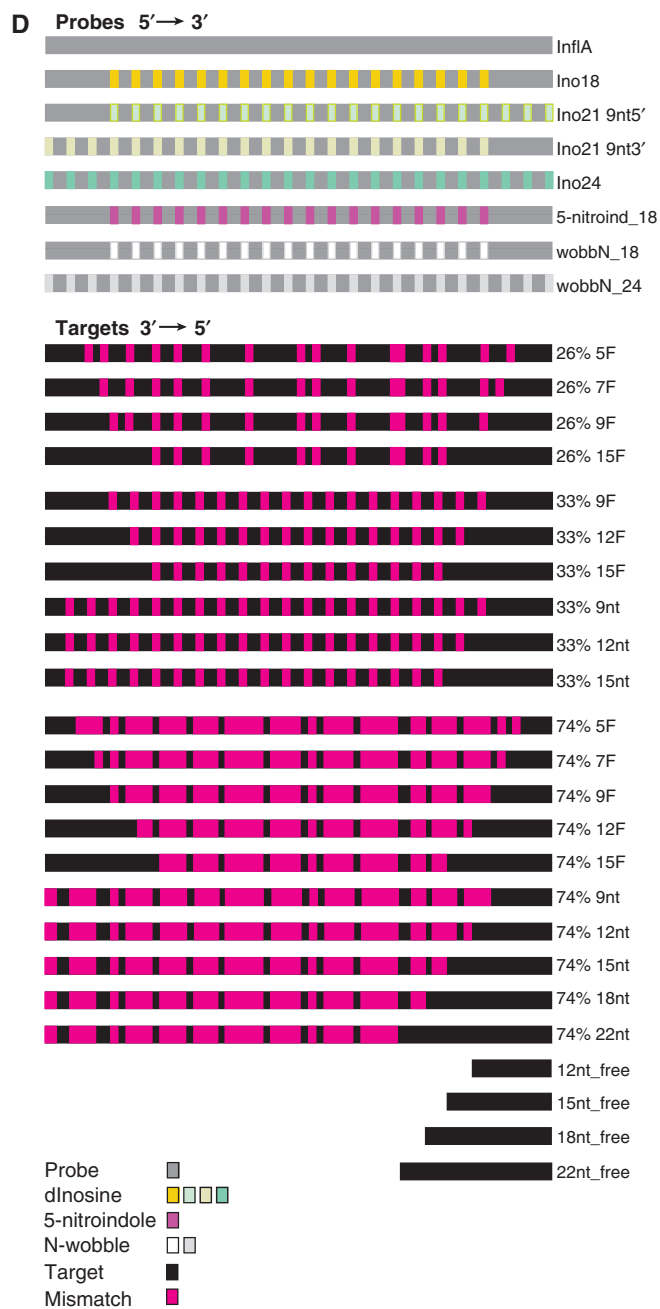


Figure 3. Continued.

region, hybridized at 45°C (43.1% of the total MFI, 2548 MFI) but not at 55°C (3.2% of the total MFI, 165 MFI), while the 74%15F target, with nine matches dispersed in the central region, hybridized at both temperatures (77.7% of the total MFI, 4597 MFI, 45°C, and 55% of the total MFI, 2768 MFI, 55°C). Thus, the perfectly matching 5' and 3' flanking regions were long enough to explain nearly all the hybridization in 74%12F and 74%15F. At 45°C, even a single contiguous stretch of 15 nt (the 74%15nt target, 17.4% of the total MFI, 1029 MFI) was enough to give a hybridization signal while, at 55°C, the long perfectly matching segment consisted of 18 nt (74%18nt target, 31% of the total

Table 2. The number of matching regions during hybridization of the InflA probe and the targets with either 14 mismatches (the 14 pm, 26%9F and 33%15F targets) or 16 mismatches (the 16 pm, 26%5F and 33%15F targets)

Target	14 mismatches			16 mismatches		
	14pm	26%9F	33%15F	16pm	26%5F	33%12F
45°C MFI% ^a	26%	41%	50%	6.1%	24%	44%
55°C MFI% ^a	4.4%	24%	39%	0.5%	5.6%	4.9%

No. of nt in each matching region	No. of matching regions			No. of matching regions		
15	–	–	2	–	–	–
12	–	–	–	–	–	2
9	–	2	–	–	–	–
6	1	1	–	1	1	–
5	5	3	–	3	5	–
4	2	1	–	2	1	–
3	3	2	–	3	4	–
2	4	2	–	8	2	15
1	–	3	13	–	3	–

^a% MFI (InflA probe against target/InflA probe against InflA target).

MFI, 2712 MFI). Thus, the perfectly matching regions have to be a specific length to overcome a high degree of variation in the remaining sequence.

These experiments, taken in conjunction with the InflA probe hybridizing to the 21-gm but not to the 21-pm target, confirm that both the number of mismatches and their distribution are important. It is reasonable to assume that perfectly matching sequences of a minimum length function as nucleation sites (15) which initiate hybridization between the probe and the target.

Long non-matching ends destabilize duplex formation

Hybridization of the InflA probe with targets containing 74% mismatching 70-mers with one perfectly matching end of various lengths (74%*x*nt) was compared with hybridization of the InflA probe with short, perfectly matching targets (12–22 nt_{free}; Figure 3A, B and D; nt sequences in Supplementary Table S2B) to analyze the effect of long mismatching ends. Utilizing the InflA probe / InflA target as reference, the 12nt_{free} target did not hybridize at 45°C (1.8% of the total MFI, 106 MFI) but the 15-, 18- and 22-nt_{free} targets gave successively stronger MFI signals (62, 98 and 123% of the total MFI; 3669, 5770 and 7269 MFI, respectively) than the 74%15nt, 74%18nt and 74%22nt targets (17.4, 55 and 62% of the total MFI; 1029, 3222 and 3637 MFI, respectively). Thus, the MFI is higher when the short, perfectly matching targets (*xnt_{free}*) hybridize with the InflA probe with only one long end protruding from the hybridized portion of the probe. Previous reports have demonstrated that 1–5 nt single dangling ends tend to stabilize duplex formation (23–26). This study shows that two long mismatching ends destabilize the hybridization of the matching part of the duplex. The long free mismatching ends could form intramolecular secondary structures that could have an effect on the hybridizing duplex.

Alternatively, the Brownian movements of the two long non-hybridized sections could mechanically stress the remaining base pairs.

Analyses of probe hybridization with targets containing regions of sm of various length

The tolerance of the probe against sm was tested using targets (70 nt) with every third nt harboring an sm (referring to the reading frame of the matrix 2 protein of the H3N2 Influenza A) in regions of different length. The 33%9F, 33%12F and 33%15F targets have a region containing 18, 16, or 14 sm between two flanking regions of 9, 12, or 15 perfectly matching nt, respectively. The 33%9nt, 33%12nt and 33%15nt targets have a region of 20, 19 or 18 sm in combination with one region of 9, 12 and 15 perfectly matching nt at the 5' end (Figure 3D, nt sequences in Supplementary Table 2A and B).

As demonstrated in Figure 3A and B (and confirmed in Figure 3C), the targets 33%9nt and 33%9F failed to hybridize with the InflA probe, while the 33%15F target hybridized at both hybridization temperatures (60% of the total MFI, 3517 MFI, at 45°C; 39% of the total MFI, 1934 MFI, at 55°C). The 33%12F target with its two 12 nt flanking regions only hybridized at 45°C (45% of the total MFI, 2635 MFI). The 33%15nt, with one contiguous region of 15 perfectly matching nt also hybridized only at 45°C (23% of the total MFI, 1352 MFI; MFI result taken from Figure 3A and B). Thus, for a 70-mer probe to hybridize with a target containing a relatively long stretch of sm, it must have (i) one uninterrupted perfectly matching region of at least 15 nt at 45°C, and longer than 15 nt at 55°C, or (ii) two uninterrupted matching regions of at least 12 nt at 45°C or 15 nt at 55°C.

dInosine-containing probes restore hybridization to targets containing sm

The targets containing sm were further tested against a set of dInosine-containing probes: one probe (the Ino18 probe) contained 18 dInosines matching the sm in the 33%9F target, two probes (21Ino_9nt5' and 21Ino_9nt3') contained 21 dInosines positioned as sm leaving a region of matching 9 nt at either the 5' or 3' end; and one probe (Ino24) contained 24 dInosines in every third base throughout the whole 70-mer probe.

Importantly, when dInosines matching the positions of the sm in the targets were included in the probes, the Ino18 probe hybridized with all the sm-containing targets, although with slightly varying MFI (Figure 3A, B and confirmed in C). The 33%9F target, which did not hybridize with the InflA probe, was able to hybridize with the Ino18 probe (66% of the total MFI, 3894 MFI at 45°C, Figure 3A; 55% of the total MFI, 2808 MFI at 55°C, Figure 3B). Even the 33%9–15nt targets, with two mismatching nt at the 3' end not covered by dInosines, hybridized well with the Ino18 probe (48–51% of the total MFI, 2818–2994 MFI at 45°C, Figure 3A; and 16–32% of the total MFI, 778–1463 MFI at 55°C, Figure 3B). Thus, as shown in the previous series of pm matched to dInosine, when the sequence is interrupted too frequently

by mismatches, leaving no suitable regions of perfect match, a probe with dInosines at the positions of variation will restore the hybridization by effectively creating the required longer matching region.

Introducing a dInosine in every third position throughout the Ino24 probe decreased the hybridization dramatically for all the 33% targets (6–16% of the total MFI, 361–920 MFI at 45°C; 1% of the total MFI, 53–79 MFI at 55°C). When utilizing the 21Ino_9nt5' probe with the 33%xnt targets, the MFI decreased because of the mismatches created by the two sm at the 3' end of the target (4–10% of the total MFI, 232–415 MFI at 45°C). In comparison, the 21Ino_9nt5' probe and the 33%xF targets, with no mismatch, hybridized at 26–43% of the total MFI (1692–2898 MFI at 45°C). Furthermore, the 21Ino_9nt3', which did cover the two sm at the 3' end of 33%xnt, hybridized more strongly (10–21% of total the MFI, 450–695 MFI at 45°C) than the 21Ino_9nt5' probe (4–10% of the total MFI, 232–415 MFI at 45°C). Both these results demonstrate that using the less stringent temperature (45°C) permits hybridization even when a large number of Inosines is present (21 dInosines in a 70-mer probe), as long as a matching region of at least 9 nt is formed in the hybrid.

The hybridization of a long probe containing dInosines is comparable with that of a long degenerated probe with the same number of N wobbles, under lower stringency conditions

The effects of probes with dInosine or wobbles in the same positions were also investigated in 3M TMAC. The presence of a dInosine in a specific position instead of a wobble would theoretically decrease the degeneration of the probe and subsequently increase the concentration of the particular probe variant. A probe with 21 N wobbles, wobbN_21, at the same positions as the dInosines in the Ino21 probe, was tested. The surprising result was that the probe containing N wobbles hybridized very well with the InflA target (29% of the total MFI, 2220 MFI, Figure 2C) and the 21-pm target (29% of the total MFI, 2190 MFI, Figure 2C). This is in the same range as hybridization of the Ino21 probe with the InflA (26% of the total MFI, 1976 MFI, Figure 2C) and 21-pm (11% of the total MFI, 794 MFI, Figure 2C) targets (39 and 37% of the total MFI versus the InflA and the 21-pm targets, see Figure 2A). These results also demonstrate that the wobb_N21 probe is not affected to the same extent as the Ino21 probe by increasing the hybridization temperature from 45 to 55°C (Figure 2C). The test was repeated by comparing an Ino18 with a wobbN_18 probe (Figure 3C and D). At 45°C, the Ino18 probe hybridized at least as well as the wobbN_18 probe while, at 55°C, the wobbN_18 probe hybridized better than the Ino18 probe. Interestingly, a probe containing 24 wobbles still hybridized better with all 33%xnt and 33%xF targets (30–45% of the total MFI at 45°C; 13–22% of total MFI at 55°C; Figure 3C) compared with Ino24 (6–16% of the total MFI at 45°C; 1% of the total MFI at 55°C; Figure 3A and B). Obviously, the 70-mer probes can accommodate multiple degenerate positions and still hybridize because the majority of probe molecules

will contain several long perfectly matching stretches created by chance. This is further deliberated under Discussion section.

The hybridization of a long probe containing dInosines is stronger than that of a long probe containing the same amount of 5-Nitroindole, at either high or low temperatures

5-Nitroindole is a second-generation universal base nt analog that was chosen for comparison with the first-generation dInosine with respect to hybridization properties in 3M TMAC. According to Loakes and Brown (1994), 5-nitroindole is less destabilizing than its 4- and 6-isomers (27) and than 3-nitropyrrole (9). A probe with 18 5-nitroindole residues (5-NitroInd₁₈) was designed; the nt analogs were distributed to match the pattern of the dInosines in the Ino18 probe (Figure 3D). The probes were allowed to hybridize with the InflA target and the set of targets with sm, 33%_{xF} and 33%_{xnt} (Figure 3C). At 45°C, hybridization of 5-NitroInd₁₈ with the InflA (44% of the total MFI), 33%_{xF} (23–37% of the total MFI), and 33%_{xnt} (4–7% of the total MFI) targets resulted in hybridization signals that were much lower than those seen with the Ino18/InflA (73% of the total MFI), Ino18/33%_{xF} (54–78% of the total MFI), and Ino18/33%_{xnt} (47–59% of the total MFI) probes (Figure 3C, 45°C). Increasing the temperature to 55°C destabilized the 5-NitroInd₁₈ probe even more, resulting in hybridization of only 1–8% of the total MFI. In conclusion, dInosine functions much better than 5-nitroindole as a universal nt analog, under 3M TMAC buffer conditions.

Hybridization of a probe containing dInosine is sensitive to mismatches neighboring the dInosine position, aiding specificity

It has been shown above that when the dInosine and the mismatch have the same distribution pattern, i.e. the dInosine masks the mismatch, hybridization can be restored (Figures 2A, B, 3A and B). We analyzed how many mismatches outside the rescuing position of dInosine (mismatch outside dInosine; mmoi) a probe can tolerate. Norovirus, Ino18 and InflA probes were tested against a set of targets whose sequences were designed to range successively from a Norovirus sequence to the InflA sequence, allowing different amounts of mismatch and mmoi to be analyzed. Norovirus is a highly variable, positive-sense RNA virus belonging to the Caliciviridae, which causes ‘winter vomiting disease’. The Norovirus sequence chosen (the capsid gene of the Norwalk-like virus, accession nr AY274264), after Blastn with the InflA probe sequence, has a short region of 8 nt that perfectly matches the end region of the InflA probe and has 10 dispersed matching nt (Figure 4A). The Norovirus target (70)0_{36_52}(0.8) (this code is explained below, and in the legend to Figure 4) was gradually changed to resemble the InflA target (0.8)51_{0_0}(70) by altering the central nt sequence. A set of targets was also created where the nine nt at the 5′ end were changed into an InflA sequence and the central region was gradually

changed from the Norovirus to the Influenza sequence, starting with (0.61)9_{27_42}(9.8). The targets were named according to the number of matching and mismatching nt in comparison with the three probes: (nt matching those of the Norovirus probe at 5′ and 3′) mismatching nt versus the Norovirus probe_Ino18 probe_InflA probe (nt matching those of the InflA probe at 5′ and 3′) (see Figure 4A; nt sequences in Supplementary Table S3A and B).

Two targets had 26 mismatches, with different dispersion patterns, after hybridization with the Norovirus or InflA probes; the distributions are shown in the (10.10)26_{10_26}(0.8) and (0.11)26_{10_26}(9.8) targets in the upper and lower panels of Figure 4A. The InflA probe did not hybridize with either of them, while the Noro probe hybridized weakly with both: 13% of the total MFI for the (0.11)26_{10_26}(9.8) target and 20% of the total MFI for the (10.10)26_{10_26}(0.8) target. The % MFI for the Noro probe was calculated by comparing the MFI with that of the Noro probe/Noro target hybridization. The stronger hybridization to the Noro probe than to the InflA probe can be explained by the distribution and length of the matching sequences between the 26 mismatches; the perfectly matching regions of 8 and 9 nt were not long enough to induce hybridization to the InflA probe, while one region of 11 nt (together with a 5-nt and a 6-nt region) or two longer flanking regions of 10 nt in the two targets was enough to induce hybridization with the Noro probe.

When the Ino18 probe (middle panel of Figure 4A) was used with a target with 1 mmoi [i.e. (0.10)35_{1_17}(9.8)] or even 10 mmoi [(10.10)26_{10_26}(0.8)] located at the 5′ and 3′ ends of the target, outside the central region containing dInosines, there was no or little inhibition of hybridization at 45°C (71% of the total MFI, 4741 MFI, and 41% of the total MFI, 2773 MFI, respectively). Interestingly, when the 10 mmoi were evenly distributed within the region of 18 dInosines, as in the (0.11)26_{10_26}(9.8) target, hybridization was lost (0.9% of the total MFI, 59 MFI). The sensitivity to mmoi adjacent to dInosine was therefore investigated further. Targets with increasing numbers (2, 4 or 5) of mmoi neighboring the positions of dInosines successively reduced the hybridization signals: 2319 MFI (35% of the total) for the (10.10)24_{12_28}(0.8) target, 927 MFI (13%) for the (10.10)22_{14_30}(0.8) target, and 375 MFI (5.4%) for the (10.10)21_{15_31}(0.8) target, all at 45°C. Thus, dInosine is sensitive to neighboring mismatches. A comparison of the sensitivity of a dInosine-free probe (InflA) and a probe containing dInosine (Ino18) to neighboring mismatches showed that the dInosine-free InflA probe can hybridize to a target with 17 evenly distributed mismatches between two perfectly matching flanking sections of 9 and 8 nt, respectively [(0.10)35_{1_17}(9.8) target, 24% of the total MFI, 1614 MFI at 45°C]. However, 7 mmoi adjacent to the dInosines completely destroyed hybridization between the Ino18 probe and the (10.10)19_{17_33}(0.8) target: 0.5% of the total MFI, 34 MFI, at 45°C. Results for 55°C are shown in Supplementary Figure S4B and Table 3. The Ino18 probe failed to hybridize when 2 mmoi were placed next to the dInosines

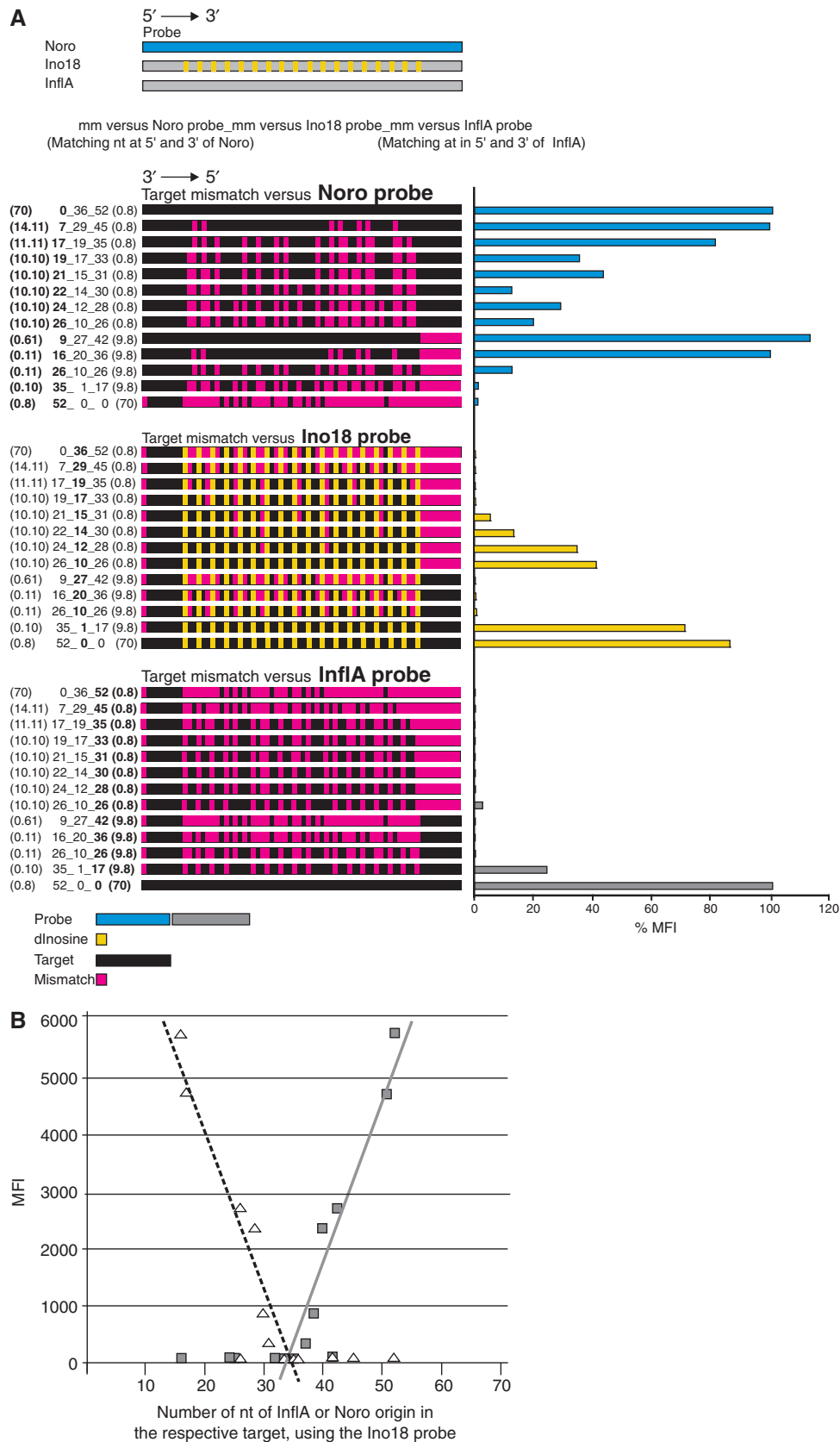


Figure 4. Analysis of specificity of hybridization by comparing a Norovirus probe, an Influenza A virus probe containing 18 dInosine, and an InflA probe. Targets containing different amounts and distributions of both Noro and InflA sequences were used. The MFI was measured from a microsphere mixture containing 2500 microspheres of each of the three probe types, the Noro, Ino18 and InflA probes, after hybridization with 0.2nM of the synthetic biotinylated target. Hybridization was performed at 45°C and 55°C, shown in Supplementary Figure S1A and B. Each

[(10.10)24_12_28(0.8); 3.1% of the total MFI, 192 MFI at 55°C]. Thus, the hybridization capacity of a dInosine-containing sequence is severely reduced when the mismatch is adjacent to the dInosine.

Figure 4A also shows how introduction of many dInosines affects the specificity. There was no hybridization between the influenza probe Ino18 with its 18 dInosines and the Norovirus target (70)0_36_52(0.8). Furthermore, although the dInosines mask 16 mismatches in the (0.11)26_10_26(9.8) target, the 10 mmoi that are in close proximity to the dInosines abolish hybridization (0.9% of total MFI, 59 MFI). In contrast, the other target containing 26 mismatching nt and 10 distant mmoi, (10.10)26_10_26(0.8), did hybridize to the Ino18 probe (41% of total MFI, 2773 MFI). Thus, the cross hybridization of a foreign (unrelated) dInosine-containing probe is dependent to a certain extent on the amount of mismatch, but is even more dependent on the distribution of mmoi.

Figure 4B and Supplementary Table S6 show the origin of the nt that are not covered by the dInosines when using the Ino18 probe (MFI values from Figure 4A). They demonstrate the number of InflA-matching nt outside the dInosine position (moi) needed for hybridization and the number of Norovirus moi causing cross hybridization. At least 37–38 InflA moi were needed to induce hybridization with the Ino18 probe but, as mentioned above, the distribution is at least as important as the actual number of matching and mismatching nt. Of the fewer than 30–31 nt that were of Norovirus origin in a target that hybridized with the Ino18 probe, 16 nt were common to both Norovirus and Influenza virus. If more than 31 nt were of Norovirus origin, hybridization to the Influenza Ino18 probe failed.

Detection of rotavirus from clinical samples using dInosine- and degeneration-containing probes

The region chosen for the 77-nt rotavirus probe, positions 112–178 in the alignment, was analyzed using the haplotype function of ConSort©, which decomposes highly variable stretches into a small number of less variable stretches (haplotypes). This resulted in two major haplotypes and probes which potentially covered all rotavirus group A variations recorded in GenBank. The two haplotype probes contained 14 and 8 dInosines, respectively, in combination with four degenerations. They were called Ino14_w4 and Ino8_w4 (Figure 5C, Inosine as yellow

and wobbles as light grey boxes). Two additional probes with fewer dInosines and more degenerations were also created: Ino11_w7 and Ino5_w7 (Figure 5C). The consensus sequence and the pattern of variation of the region chosen for the probe are shown in Figure 5A. The sequences of the four probes are shown in Supplementary Table S4A and B and (schematically) in Figure 5C.

The degenerated LNA-containing fw-primer and the long degenerated dInosine-containing biotinylated rev-primer generated a single band of the correct size, 502 nt, when analyzed by electrophoresis using EtBr-stained agarose gel in all five clinical samples (data not shown). All four microsphere-bound probes detected the consensus synthetic rotavirus target as well as the amplified rotavirus nucleic acid from all five clinical samples. Interestingly, the four probes hybridized almost equally well within each sample. It was found that an asymmetric PCR of the clinical samples was necessary in order to obtain an MFI of reasonable strength from the probes (data not shown). This was probably because the complementary strand of the PCR product outcompeted the probe due to an affinity between the two strands that was higher than that between a dInosine-containing degenerated probe and the target strand. The data in Figure 5B are from one of the experiments using samples run in duplicate. The PCR products were sequenced (Supplementary Table S4), revealing that the Ino14_w4 and Ino11_w7 probes, which belonged to the same haplotype, covered the variations in all positions in all samples. However, the other pair of probes (Ino8_w4 and Ino5_w7) had one mismatch against clinical samples 1 and 2 and two mismatches against clinical samples 3, 4 and 5, as well as the consensus synthetic rotavirus target (magenta colored boxes in targets in Figure 5C, Supplementary Table S4). In conclusion, the long dInosine-containing degenerated probes worked well as variation-tolerant probes, covering variations, accepting a few mismatches, and still remaining specific (neither of the Rotavirus probes hybridized with the InflA target).

Development of scores for local and distant cooperative interactions

Once we had these experimental data, we tried to develop a unifying view of them. The ΔG predicted by the Visual OmpTM 7.0 software was compared with the percentage of the total MFI for each probe and target combination, including dInosine-containing probes (Figure 2A, 3A

Figure 4. Continued

experiment contained triplet samples and was repeated two or three times. The Noro target, complementary to the Noro probe, was mutated gradually towards the InflA target sequence, resulting in twelve new targets and the InflA target (Figure 4A). The sequences of the panel of probes and targets are displayed in Supplementary Table 3A and B. The system used to name the targets is based on matching and mismatching the nt in the target to those in the Norovirus probe, the Ino18 probe, and the InflA probe: (nt matching those of the Norovirus probe at 5' and 3') mismatching (mm) nt versus the Norovirus_Ino18_InflA probe (nt matching those of the InflA probe at 5' and 3'). (A) Hybridization of three probes, the Norovirus (blue), InflA (grey), and Ino18 (yellow/grey), showing the distribution of yellow dInosines in the InflA sequence) probes, with a panel of gradually changing Influenza A and Norovirus targets. In the panel of thirteen target molecules (black), the Noro target sequence (complementary to the Noro probe) is at the top, with gradual change towards the InflA target (matching the InflA probe) at the bottom. The target panels show the patterns of mismatches (magenta) in each target versus the Noro probe (upper panel), the Ino18 probe (middle panel) and the InflA probe (lower panel) and the respective percentages of the total MFI, representing the extent of hybridization between the Noro, Ino18 and InflA probes and the targets at 45°C. (B) The x-axis of Figure 4B shows the total number of nt in each target that originated from either Noro (white triangles) or InflA (grey squares) and that are outside the position of the 18 dInosines when hybridized with the Ino18 probe, compared with the MFI (from Supplementary Figure S1A) for each combination.

Table 3. Summary of the effect of the number of mismatched nt outside the position of dInosine (mmoi)

Probe	Ino18			
	Nr of mmoi within 18sm region, %MFI ^a			
	45°C		55°C	
Target	Hybr	No hybr	Hybr	No hybr
(70) 0_36_52 (0.8)		26 mmoi, 0.4%		26 mmoi, 0.7%
(1.14) 7_29_45 (0.8)		19 mmoi, 0.5%		19 mmoi, 0.5%
(11.11) 17_19_35 (0.8)		9 mmoi, 0.4%		9 mmoi, 0.6%
(10.10) 19_17_33 (0.8)		7 mmoi, 0.5%		7 mmoi, 0.4%
(10.10) 21_15_31 (0.8)	5 mmoi, 5.4%			5 mmoi, 0.5%
(10.10) 22_14_30 (0.8)	4 mmoi, 13%			4 mmoi, 0.6%
(10.10) 24_12_28 (0.8)	2 mmoi, 35%			2 mmoi, 3.1%
(10.10) 26_10_26 (0.8)	0 mmoi, 41%		0 mmoi, 18.2%	
(0.61) 9_27_42 (9.8)		26 mmoi, 0.4%		26 mmoi, 0.6%
(0.11) 16_20_36 (9.8)		19 mmoi, 0.6%		19 mmoi, 0.6%
(0.11) 26_10_26 (9.8)		9 mmoi, 0.9%		9 mmoi, 0.5%
(0.10) 35_1_17 (9.8)	0 mmoi, 71%		0 mmoi, 40%	
(0.8) 52_0_0 (70)	0 mmoi, 86%		0 mmoi, 59%	

The number of mmoi affected the outcome of hybridization (hydr) between the Ino18 probe and the respective target at 45°C and 55°C.

^a%MFI=MFI of Ino18 probe and respective target / MFI of InflA probe and InflA target.

and 4A). For the sake of simplicity, data from 5-nitroindole and N-wobble-containing probes were omitted. The results demonstrated that some probe/target combinations that hybridized well in practice, had very low predicted ΔG values (Figure 6B), e.g. the InflA probe hybridizing to the 74%12F (43% of the total MFI, $\Delta G = -20.62$), 74%15F (77% of the total MFI, $\Delta G = -26.7$), 74%18 nt (55% of the total MFI, $\Delta G = -29.24$), and 74%22 nt (61% of the total MFI, $\Delta G = -34.55$) targets.

When the results of the new NucZip scoring system were scored against the %MFI in Figure 6A, which shows all the target and probe combinations plotted in Figure 6B, it was found that they were more highly correlated with the experimental data than the predicted ΔG . To investigate these differences, each outlier in Figure 6B was connected to its plot position in Figure 6A; see Figure 7A and B. Figure 7 shows that probe–target combinations containing many mismatches and dInosines were the main causes of the lower correlation between predicted and observed hybridization in Figure 6B. However, hybridizations between a long probe and a short target were not included in Figures 6 and 7. Nor were data from probes containing 5-nitroindole or N-wobbles, because a full investigation such as this would require many more observations and would be out of the scope of this article. The NucZip results are further discussed in the Discussion section.

Thus, when the actual degree of hybridization for the entire data set (265 probe–target combinations) was matched with the predicted ΔG in Visual OmpTM 7.0, an only moderately precise correlation was obtained. The hybridization of combinations involving many mismatches and many dInosines was poorly predicted. However, when the NucZip algorithm was used, a

higher degree of correlation was observed. The adjusted determination coefficient (R_a^2) was 0.8636, indicating that 87% of the variation was explained by the NucZip algorithm, while the best fit of MFI% to the Visual OmpTM 7.0 predictions gave a determination coefficient of 0.7505, indicating that 75% of the variation was explained by NN theory (as embodied in Visual OmpTM 7.0). NN theory was thus insufficient for predicting hybridization under the hybridization conditions of our study. A high number of mismatches and dInosines gave hybridization predictions in Visual OmpTM 7.0 that were too low (Figure 6B). The NucZip algorithm, which takes into account the length of matching segments and cooperativity effects within and between matching oligonucleotide segments, increased the accuracy of hybridization prediction. dInosines were scored intermediate between matches and mismatches.

Other hybridization prediction algorithms are available on the Internet. However, when we compared the delta G predictions obtained from IDT Oligo Analyzer (<http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/>), which uses a proprietary algorithm, with our experimental data, the correlation was poor (Supplementary Figure 2). Although the exact experimental conditions (3M TMAC and 45°C) were not represented, this is not likely to have caused the low correlation.

DISCUSSION

Nucleic acid hybridization is fundamental to many molecular biology applications, and is expected to grow in significance as nanomedicine joins molecular medicine at the cutting edge of research (28). In particular, biomedical applications of hybridization such as detection of variable viral target sequences are highly dependent on a

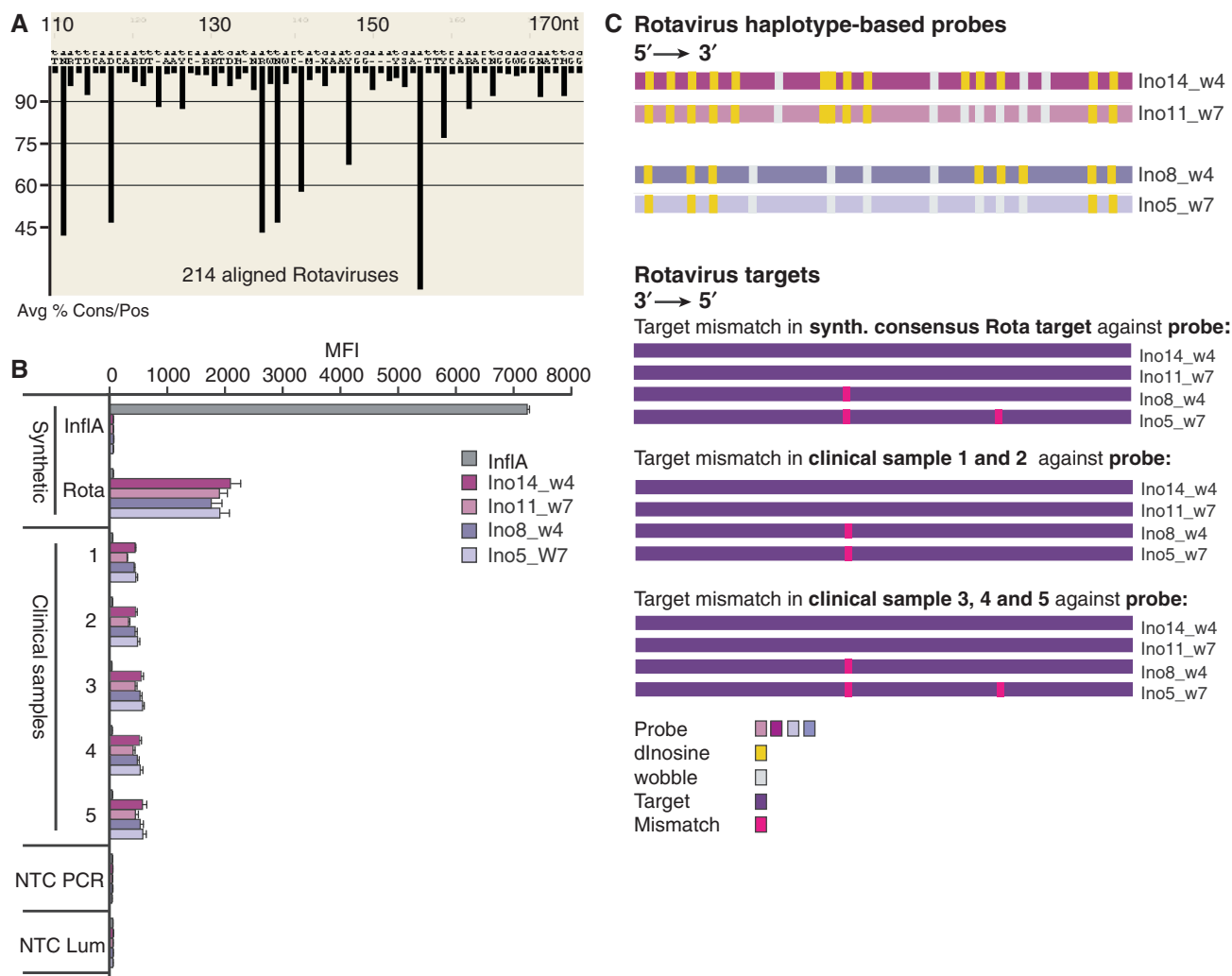


Figure 5. Detection of Rotavirus in clinical samples. (A) The variations in Rotavirus were mapped using BLASTn (16) and ConSert©, to find suitable regions for primers to be used in reverse-transcription PCR and a conserved region for a detection probe. The length of the black bars represents the frequency of variation as an average percentage conservation at each nt position (y-axis). The figure shows the alignment and variation of 214 Rotavirus A sequences in the nt position of the probes displayed in Figure 5C. ConSert© was used to group the variations of the probe region into two haplotypes, which were then used to construct the probes shown in Figure 5C and Supplementary Table S4A and B. (B) Nucleic acid from clinical samples was extracted, amplified and biotinylated with an asymmetric RT-PCR, and 5 µl was incubated with 2500 beads of each of the four types of Rotavirus probe and InflA probe. As positive controls for the hybridization, 0.2 nM of single-stranded biotinylated InflA and Rotavirus were used. The samples were tested as double samples. (C) A schematic illustration of the positions of dInosines (yellow boxes) and wobbles (Y, R, W, D) in the Rotavirus probes, and the mismatching nucleotides (magenta boxes) in the synthetic target and the clinical samples revealed after sequencing of the PCR products.

precise understanding of the process involved. A probe that has a broad detection spectrum should be as specific as current narrower probes while retaining the ability to cover the biologically or clinically relevant sequence variants of specific microbes. The design of long mismatch-tolerant probes demands knowledge about hybridization in the presence of mismatches, degeneracy and nt analogs.

In pursuit of this level of understanding, and in order to obtain reliable hybridization data, we chose to use the Luminex suspension array system in our studies. The inherent ability of the system to report the median of a high number of measurements (i.e. measuring hybridizations signals from 100 different beads) provides highly reliable data. Moreover, hybridization equilibrium is

reached more rapidly using the suspension array system (taking around 15 min) than by solid phase hybridizations such as micro arrays (often overnight). A probe length of 70 nt was selected for our studies because of the elevated mismatch tolerance of this length compared with shorter probes (29). However, the advantage of the extended length of the probe could possibly be countered by a loss of specificity. Hybridization studies using long (50 or more nt) probes in a 3MTMAC buffer system have not been reported previously; in reports using other hybridization systems, however, it has been suggested that 50-mer or 70-mer probes should contain no more than 15–20 contiguous nt complementary to non-targets (29,30). Previously, the hybridization properties of long probes have been analyzed using microarrays, with overnight

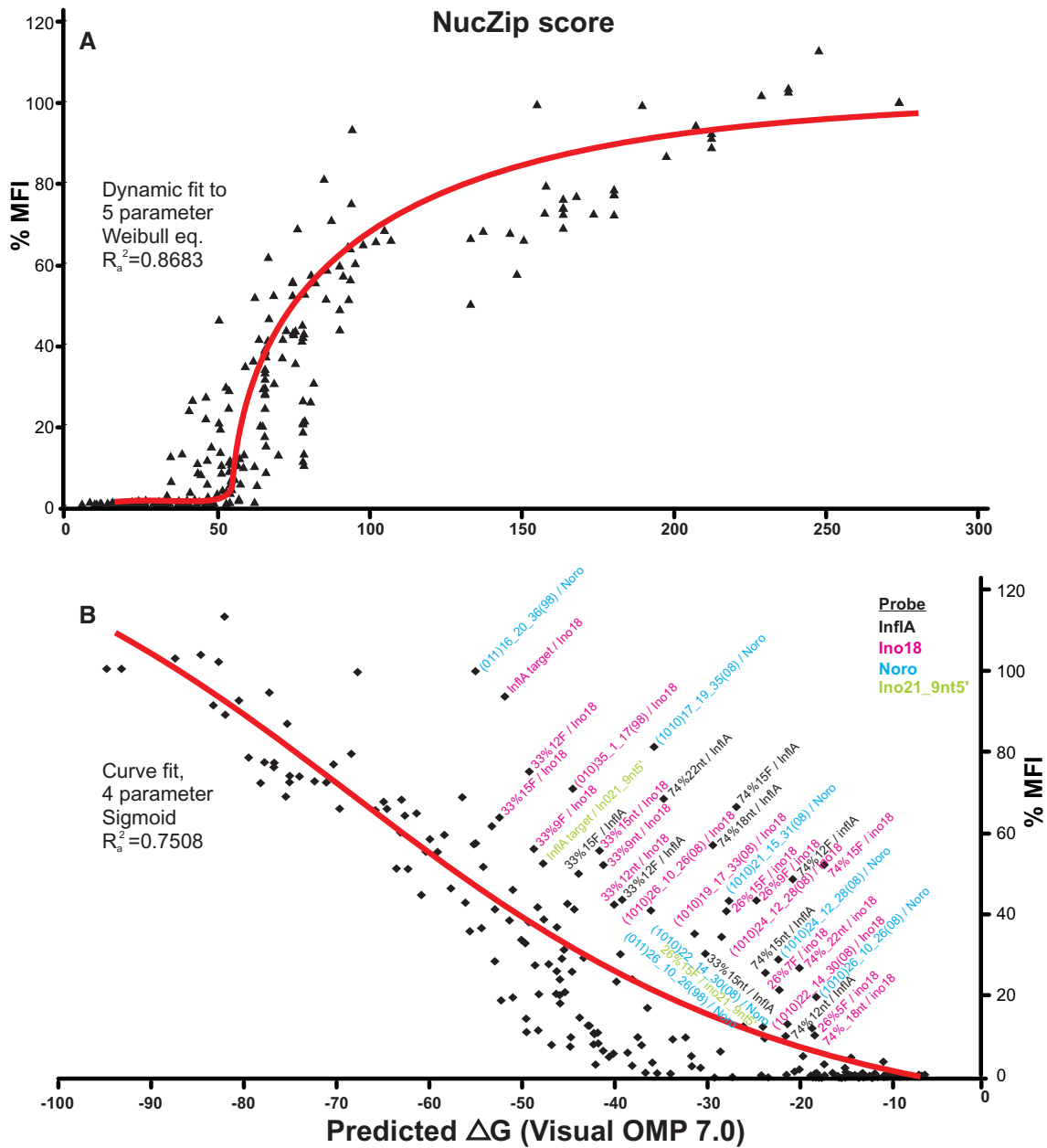


Figure 6. Prediction of hybridization using two algorithms. (A) NucZip score versus %MFI. The results of the NucZip scoring system versus the %MFI for the same target and probe combinations that were used for Figure 5A. A five parameter Weibull function gave the highest correlation [$f = \text{if}(x, x_0 - b * \ln(2)^{(1/c)}; y_0; y_0 + a * (1 - \exp(-(\text{abs}(x - x_0) + b * \ln(2)^{(1/c)})/b)^c))$], where $a = -274.1515$, $b = 1288.3093$, $c = 0.4614$, $x_0 = -639.0365$ and $y_0 = 4.7970$; $R = 0.9306$ and $R_a^2 = 0.8636$]. (B) Predicted ΔG versus observed hybridization (% MFI). The percentage MFI from all hybridizations (at 45°C) of the 70-mer probes and 70-mer targets from the three panels presented in Figure 2A, 3A and 4A were plotted against the predicted ΔG calculated in Visual OMP™ 7.0 (DNA Software). The predicted ΔG was obtained for the interaction between probes and targets in 3M TMAC buffer at a hybridization temperature of 45°C. The percentage MFI is the MFI signal of a probe hybridized with a study target divided by the MFI signal of the same probe hybridized with its perfectly matching target (e.g. InflA probe against InflA target) at the same temperature. Regression lines were calculated using the SigmaPlot dynamic curve fitting system. A five parameter sigmoidal function gave the highest correlation [$f = y_0 + a / (1 + \exp(-(x - x_0)/b))$], where $a = 122.5959$, $b = -9.5493$, $c = 0.2793$, $x_0 = -78.0478$, $y_0 = -14.5630$; $R = 0.8689$, and $R_a^2 = 0.7505$].

hybridization in buffers containing formamide/SSC (sodium chloride/sodium citrate) at 42–53°C (30–34). Letowski *et al.* found that, under microarray conditions, mismatches grouped at the 5' or 3' end of a 50-mer probe affected the binding to a target less than if the mismatches were distributed throughout or centered in the probe. Furthermore, the 50-mer probes with mismatches

distributed along the whole probe were more destabilized than the probes that had mismatches centered in the duplex (34). When Deng *et al.* studied mismatches in 50-mer microarray probes with 1–7 pm in different distribution patterns, they concluded that the signal intensity was decreased more by evenly distributed than by randomly distributed pm (32). When 60-mer

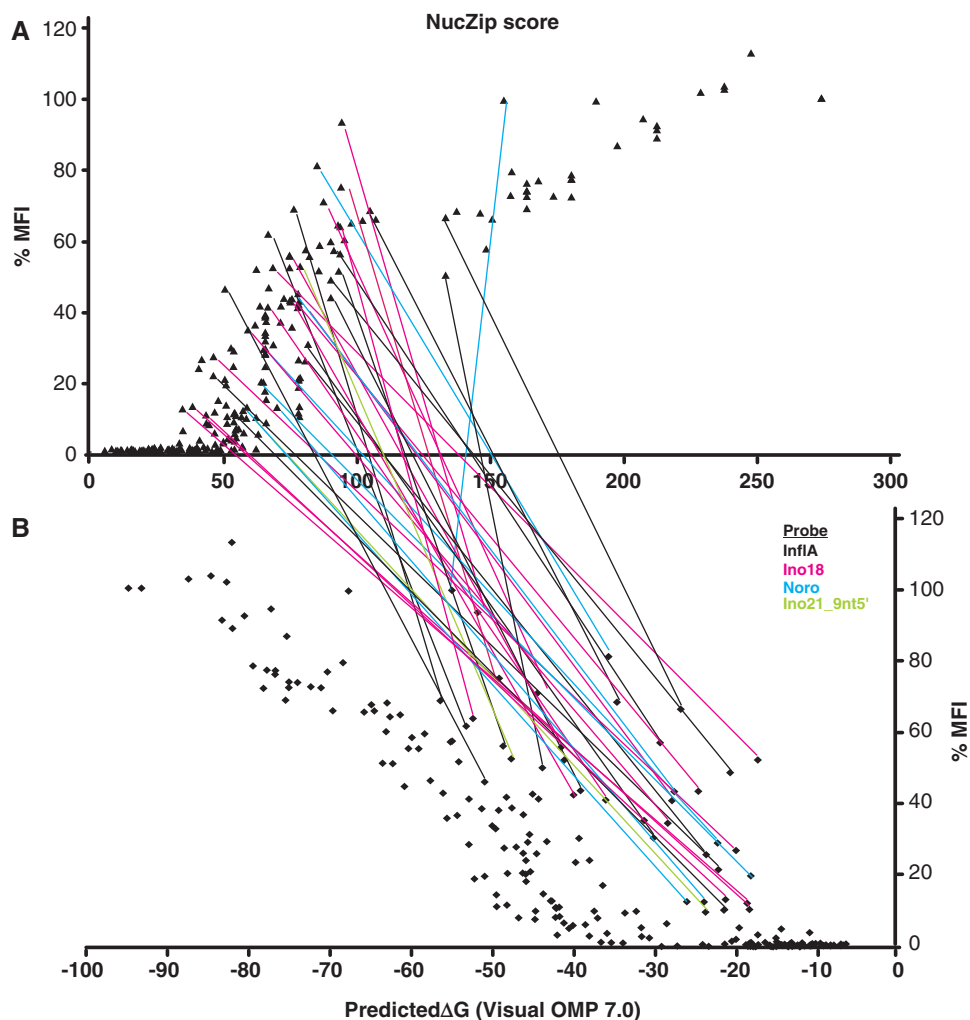


Figure 7. The same data as in Figure 6, but with lines connecting the results for the same probe–target combinations using the two prediction algorithms. (A) NucZip scores for the probe–target combinations. (B) Visual OMP deltaG predictions for the same combinations.

oligonucleotides were hybridized in a microarray, mismatches located near the middle of the probe resulted in a greater reduction of signal intensity than those located at the ends (33). Additionally, microarray experiments with short oligonucleotides (16–40 nt) and one mismatch or a nt insertion in all positions show that the hybridization signal decreases when the mismatching portion is centered (35,36). The matching segments are shorter when mismatches are centered than when they are located peripherally. Our results, using microsphere-bound 70-mer probes and Luminex technology in a buffer containing 3 M TMAC, confirm that the distribution of the mismatches is of great importance and that hybridization is stronger when there are a few longer uninterrupted sequences than when there are many short sequences. These effects are formulated in the NucZip algorithm.

Furthermore, it has previously been shown that two different oligonucleotides of 18 nt, complementary to the inner and outer portions of a 25 nt probe, could hybridize in solution with equal efficiency but, when the probe was coupled to a solid phase via a C6 linker, the 18-nt target

complementary to the outer part of the 25-nt probe bound more efficiently than the 18-nt target complementary to the inner part, close to the solid phase (37–39), cf (33). On our probes, the 5' end was coupled to the microspheres via an amino C12 linker. We did not observe any significant differences between matching stretches close to the bead surface and far from it and conjecture that perhaps the long linker allowed for greater accessibility.

It is possible to create a probe against a target with high nt variation, such as an RNA virus, by using degenerated bases at variable positions but the degeneracy of the probe is dramatically increased by each wobbling base, thus decreasing the effective probe concentration. For instance, a sequence with two wobbling bases present in nine positions of variation would give a degeneracy of 512 unique sequence combinations, while a target with 14 variations including A, T, C or G would demand a set of 268×10^6 unique probes (degeneracy 268×10^6). Honoré *et al.* successfully used 18–23 nt probes with a degeneracy of up to 512 in 3 M TMAC buffer (40). In the more stringent PCR buffers, the usage of probes with a degeneracy greater than 10 is not often reported (41).

Degenerated primers have the property of being ‘forgiving’ (41–43). This is because the amplicon from a previously successful primer is a target for the same pool of primers in the next round of amplification, leading to an accumulation of amplifiable targets. However, the situation for a probe is different. A degenerate probe will always face the same target variation. Therefore, universal bases like dInosine may be more useful than degenerated sites for probes, as long as the hybridization strength (represented by T_m or $-\Delta G$) is good enough. The introduction into a probe of a universal base like dInosine instead of a wobbling base reduces the complexity of the oligonucleotide mixture and increases the actual number of hybridizing oligonucleotides.

Previously, Honoré *et al.* introduced up to three dInosines in radioactively labelled short oligonucleotide probes, 18–23 nt, in dot-blot hybridization, using a buffer containing 3 M TMAC (40). The aim was to reduce the degeneracy in probes used for screening cDNA libraries. They found that dInosines had a slightly destabilizing effect on hybridization, especially when hybridizing against A, G and T, but that this could be minimized by reducing the hybridization temperature. However, the behavior of dInosine in long probes in 3 M TMAC has, to our knowledge, not previously been systematically explored. The 3 M TMAC is known to increase the binding contribution of the A:T base pairs, resulting in a similar contribution to T_m to that from the G:C base pairs. The high ionic strength makes this an environment of relatively low hybridization stringency. The general trend shown in our study (e.g. Figure 2A and B), that dInosine in the probe decreases hybridization in 3 M TMAC, indicates that TMAC did not enhance the binding strength of dInosine base pairs as much as that of A:T base pairs; cf (8). In a segment with dInosines at every third base, such as in the Ino18 probe, every matching nt neighbors a dInosine, i.e. it will not bind as strongly as a probe containing neighboring matches. Clearly, dInosine matches cause less destabilization than mismatches, and allow hybridization of probe/target combinations with many short matching segments, like the InflA probe / 21-mer target combination. It is reasonable to assume that when a probe fails to hybridize

due to a high number of mismatches in the target, dInosines at these positions will restore hybridization, since dInosine appears to bridge adjacent matching stretches, increasing their ability to nucleate.

Our results confirm the findings of Honoré *et al.*, despite differences in (i) methods of detection, (ii) hybridization time, and (iii) length of probes. The experience gathered in this work indicates that dInosine base pairing can be considered intermediate between a match and a mismatch, when carried out in 3 M TMAC. Furthermore, our results indicate that dInosine causes less destabilization when hybridized with a C and an A, than when hybridized with a G and a T (7,8,40), in 3 M TMAC; e.g. the Ino18 probe hybridized more strongly with the 33%9nt and 33%15nt targets than with the 33%12nt target (Figure 3C and D and Supplementary Table S5). To lessen this effect and to be able to use the same hybridization temperature for a panel of probes containing no or different amounts of dInosine, it is preferable for the probes to be long, like the 70-mer probes investigated here.

The effects of the universal base dInosine were also compared with those of N wobbles. Thus, at the lower temperature, a dInosine-containing probe hybridized more strongly and, at the higher temperature, the N wobble probe hybridized more strongly. To understand how the highly degenerated probes, wobbN_21, wobbN_18 and wobbN_24, hybridized so well, we calculated the probability of randomly achieving an extension of the matching regions at the 5' and 3' ends of the wobbN_18 probe (Table 4). The probability that the closest N wobble to either the 5' or 3' end would be a perfect match is 0.5. Thus, 50% of the pool of degenerated probes have a 3 nt longer perfect match (12+9 or 9+12 matching nt at the 5' and 3' flanking regions) which, according to our results with non-degenerated probe/target combinations, should lead to rather good hybridization of the wobbN_18. In fact, wobbN_18 hybridization was similar in strength to that of the InflA probe to the 33%12F or 33%15nt targets. Furthermore, the probability of having several additional 5-nt matching regions in the central region is also high, probably giving rise to many more combinations in the same pool that matched

Table 4. The probability of getting matching region/s of a certain length in the outer segments of the wobbN_18 probe

Matching nt in flanking regions	Sequence												Probability ^a
9nt + 9nt	xxx	xxx	xxx	Nxx	Nxx	Nxx	(Nxx) ₁₂	Nxx	Nxx	Nxx	xxx	xxxx	
12nt + 9nt				1/4	–	–	^a	–	–	–			
9nt + 12nt				–	–	–	^a	–	–	1/4			0.25 × 2 = 0.5
15nt + 9nt				1/4	1/4	–	^a	–	–	–			
9nt + 15nt				–	–	–	^a	–	1/4	1/4			(0.25 ²) × 2 = 0.125
18nt + 9nt				1/4	1/4	1/4	^a	–	–	–			
9nt + 18nt				–	–	–	^a	1/4	1/4	1/4			(0.25 ³) × 2 = 0.03125
12nt + 12nt				1/4	–	–	^a	–	–	1/4			0.25 ² = 0.0625

The sequences of the wobbN_18 probe are displayed with x as the nt matching the target sequence and N as the wobbling nt A, C, G or T. The probability of increasing the length of matching flanking 5' and 3' regions of the probe is calculated for a few examples.

^aThere is a high probability of several additional matching regions of 5 nt in the central region, which will contribute to hybridization. x, matching nt; N, wobble of A, C, G, or T.

even better. By restricting the wobbles to 3 (e.g. a D or a B) or 2 (e.g. a Y or a T) nt, the probability of a match becomes even greater. Thus, in a highly degenerated probe with at least one continuous region of perfectly matching nt, a large part of the pool will extend this region and contribute to nucleation, zipping and hybridization. The behavior of the highly degenerated probes is encouraging and in accordance with the NucZip model, which predicts that the high likelihood of several matching stretches of 5 nt or longer will result in significant hybridization.

One of the aims of this study was to investigate the binding capacity of dInosine in 3M TMAC. 5-NitroIndole was chosen as a comparative universal base. The results shown in Figure 3C demonstrate that 5-NitroIndole in the 5-NitroInd₁₈ probe had a much greater destabilizing effect than dInosine in the Ino₁₈ probe, without the same capacity to rescue hybridization with a target containing many sm. Furthermore, the 5-NitroInd₁₈ probe was more affected than the Ino₁₈ probe when the hybridization temperature was raised from 45 to 55°C. It is concluded that, under 3M TMAC buffer conditions, dInosine is a better choice than 5-NitroIndole when designing a variation-tolerant probe with as little degeneration as possible.

Recently, Majlessi *et al.* (15) studied the nucleation process during double helix formation of short probes, 18–28 nt, with RNA or DNA targets of varying lengths and number of mismatches, in a buffer containing lithium succinate and lithium lauryl sulfate at pH 5.1. Hybridization is initiated by random collisions, but occasionally the complex is stable enough to nucleate the hybridization process. After investigating their model, they suggested that one nucleation region of 9 nt is not enough for further zipping and formation of a double helix. Instead, the first nucleation site needs a second nucleation site so that they can then cooperatively induce the zipping mechanism. They reported that inactivation of one of the 9-nt sites reduced hybridization >2-fold. Interestingly, one complete turn of a dsDNA molecule consists of 10.4 nt (44). It is conceivable that the first (often temporary) contact between two single nucleic acid strands (nucleation) should not exceed one turn of the dsDNA helix in length, to avoid torsional disturbance and faulty interlocking of the strands. From this point of view, the nucleation site should be long enough to minimize false contacts, and short enough to have minimal steric effects on the strands. Nucleation sites of 6–9 nt fulfil these criteria. The chance of two random single strands matching at a hexanucleotide is 1/16394, and at a nonanucleotide is 1/1048576. A matching nonanucleotide will thus reduce the ratio of random successful to unsuccessful nucleations, i.e. those which do not lead to further hybridization in the subsequent zipping phase, a million-fold. The subject is far beyond the scope of this article; however, our data, using 70-mer probes in 3M TMAC buffer, reveal that a target with two separate regions of 9 nt was enough for efficient hybridization when several shorter regions of 5–6 nt were available between mismatches during the hybridization process (InflA probe/26%9F, 2583 MFI). Increasing the number of mismatches, i.e. shortening the matching regions between the 9 nt flanking regions, caused

failure of hybridization (InflA probe/33%9F, 133 MFI and InflA probe/74%9F target, 86 MFI). In contrast, the Ino₁₈ probe, with dInosines covering the mismatches and nine matching nt at the 5' and 3' ends, hybridized strongly (Ino₁₈/33%9F, 3894 MFI). Furthermore, the InflA probe/74%15nt or InflA probe/74%18nt combinations showed that one region of 15–18 nt was enough to induce and sustain hybridization reproducibly, even if the rest of the 70-mer probe contained 74% mismatches. Having two perfectly matching regions of 15 nt (15F) compared to one region of 15 nt (15nt) gave 2.6-fold higher MFI for 33%15F compared to 33%15nt and 4.5-fold higher MFI for 74%15F compared to 74%15nt, at 45°C. The Ino₂₄ probe, with no dInosine-free matching trimers, hybridized inefficiently with all targets (361–920 MFI at 45°C), showing that dInosine is relatively inefficient in creating nucleating regions. Thus, two nucleation sites of 9 nt are enough to cause hybridization in 3M TMAC when they are placed next to each other to form a longer region or when there are enough shorter matching regions of 5–6 nt between them. Alternatively, dInosines could bridge the mismatches between the 9-nt regions.

Earlier work has indicated that a region of 15 nt in a 50-mer probe or 20 nt in a 70-mer probe could cause significant cross hybridization in microarray hybridization experiments (30,31) and our data agree with this. A 70-mer probe is able to hybridize with a region of 15 nt in an otherwise highly mismatching target in 3M TMAC (74%15nt, Figure 3A). To summarize, the current study shows that a dInosine-free probe of 70 nt needs (i) at least three regions of at least six perfectly matching nt, (ii) two stretches of 12–15 perfectly matching nt, or (iii) one stretch of 15–18 perfectly matching nt to result in measurable hybridization. Probes with a high number of dInosines positioned at sites of variation need shorter matching regions than dInosine-free probes. It is suggested that this is probably because dInosine participates in nucleation and zipping during the hybridization process. Thus, a probe with 18 dInosines which match mismatches in the target needs either (i) two regions of 9 nt if the hybridization temperature is 55°C, or (ii) one region of nine perfectly matching nt at 45°C.

As also shown in the study, the risk of cross hybridization when using an nt analog like dInosine is minimal, since dInosine is sensitive to a mismatch in the position next to it and >5 mmoi will reduce hybridization. On the other hand, one should be aware that if an unintended target has many mismatches covered by dInosine and only a limited number of mmoi (<5 mmoi), this could lead to cross hybridization and false positivity. Furthermore, the assumption that sm all differ at the third codon base is an oversimplification. Some synonymous codons also differ at the first and second bases. Thus, even if the Ino₁₈ probe could hybridize when 11 of 17 trimers were intact, with perfect matches at bases 1 and 2, the tolerance to mmoi of a highly dInosine-substituted probe like Ino₁₈ is limited. Around half of its six surplus trimers must be reserved for sm occurring at codon positions 1 and 2. This leaves three trimers available for non-synonymous mismatches. However, a long probe

is more likely than a short probe to have matches not neighboring mismatches.

The NN theory was developed for hybridization of short oligonucleotides in solution (45). Its application to surface-bound oligonucleotides has not been precisely studied. Hooybergs *et al.* studied hybridization of 30-mer surface-bound oligonucleotides with a 20-mer linker to 30-mer targets in solution, with no, one or two evenly spaced mismatches (46). Although NN theory was approximately corroborated, NN factors had to be recalculated to give an approximate fit to experimental data. Moreover, the adsorptive (Langmuir) behavior deviated from expectation at high target concentrations. Thus, many unresolved questions regarding hybridization behavior remain.

The concept of NucZip, with both local and distal cooperativity contributions, is an attempt to predict the hybridization behavior of most 70-mer probe–target combinations under the given conditions (Figures 6A and 8). The NucZip algorithm is now under revision to include probes containing 1, 2, 3 or 4 nt wobbles, as well as taking into account the results with the universal base 5-nitroindole. The algorithms, schematically described in Figure 8, (i) were based on our experimental data (described above) and (ii) included highly matching nucleation sites extending beyond the neighboring nt. Thus, unlike the NN theory, NucZip takes the effects of matches at longer distances into account. The unique

property of 3 M TMAC to provide a roughly equal contribution to hybridization by A:T and G:C pairs justifies simple computational approaches. The well known additivity of binding contributions per nt inherent in ΔG calculations according to the NN theory (22,45) indicates that hybridization can be treated in a relatively simplistic way. Our concepts were based on the finding that the longer the sequence of uninterrupted matching nt, the more stable is the hybridization. Our calculations were thus focused more on binding than on destabilization, i.e. the use of positive rather than negative contributions. One of the weaknesses of the NN theory is that it adds all contributions, positive or negative, to a grand sum. The negative contributions are subtracted for the whole molecule, instead of in the local context where they belong. In our approach, the binding is first assessed locally, mimicking nucleation, and then extended cooperatively to the whole molecule, mimicking the zipping process. It is binding that keeps the hybrid together, and it is thus logical to focus on binding.

NN theory does, to some degree, predict that the distribution of the mismatches is an important factor insofar as it affects the nearest neighbors. However, the concentration of this theory on the nearest neighbor disregards the importance of longer matching stretches. The importance of mismatch distribution and uninterrupted matches is exemplified by the stronger signal for hybridization of the InflA probe with the 21-gm target (30% of the total MFI, 2005 MFI at 45°C; 8% of the total MFI, 443 MFI at 55°C) compared to the abolished signal when using a 21-pm target. This is demonstrated in Table 2 by the InflA probe hybridizations with targets containing 14 or 16 mismatches in different distribution patterns, where long uninterrupted perfectly matching sequences favoured hybridization. The cooperativity in hybridization beyond the nearest neighbor that was noticed when two or more matching trimers neighbored each other strengthened hybridization more than when the same numbers of matching trimers were separated by mismatches. In Figure 3A (45°C), the InflA probe hybridized better with 74%12nt (10% MFI), 74%15nt (26% MFI), or 75%18nt (57% MFI) targets than with the 74%9F target, with its 9 nt+9 nt of perfect match (3% MFI). Furthermore, at the higher temperature (Figure 3B, 55°C), the 74%18nt (37% MFI) and 74%22nt (66% MFI) targets resulted in good hybridization while the 74%12F target, with its 12 nt +12 nt perfectly matching regions, failed (3% MFI).

Three matching trimers account for approximately one turn [10.4 nt (44)] of the helix of a dsDNA molecule. It is thus likely that two long DNA strands become entangled or conformationally committed when at least one turn of the helix has been completed. Nucleation has to proceed rapidly, without too much torsion and entanglement, to allow many encounters in a short time. We envision that contact between two strands extending to 9 nt allows rapid comparison between strands with only local torsional disturbance. If the brief contact does not achieve binding strength over a certain threshold (the nucleation threshold), the strands separate and new comparisons are made. If the nucleation threshold is exceeded at initial contact,

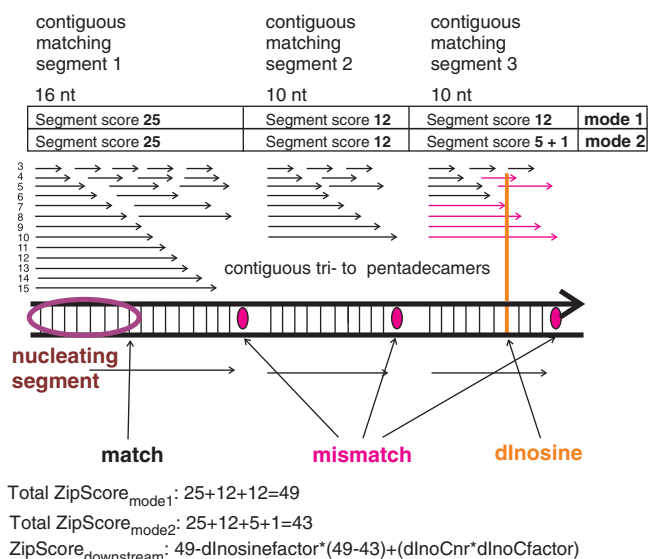


Figure 8. The zipping component of NucZip. When a 6–9-bp sequence fulfilling the nucleation criteria has been detected, hybridization up and downstream of the nucleating site is attempted (zipping). The figure shows the downstream zipping process, with successive accumulation of score within a matching segment arising from the trimers, tetramers, etc. up to pentadecamers (each scoring equally) which fit into it. In this way, a longer matching segment gets more than a linear increase in score relative to a shorter one. Zipping extends from the potential nucleation site, terminating with a mismatch or the end of one of the strands. dInosines are counted as intermediate between a match and a mismatch, as described in the ‘Materials and Methods’ section. If several consecutive matching segments are encountered, their scores are added.

binding extends further up and downstream, i.e. zipping occurs. This proceeds as long as the binding strength remains sufficient. The strands are held together chemically by base–base interactions and topologically by the multiple turns of intertwinement. It was a challenge to model this process. In the NucZip model, the program first tests for possible nucleation sites and selects the two highest scoring sites for further evaluation. Zipping is then performed up and downstream from each suggested nucleation site. The zipping algorithm symbolizes the successive cooperativity of binding by adding the number of successive tri- to pentadecamers for each matching segment, each of which ends either in a mismatch or at the end of one or two of the oligonucleotides. The contribution of added dInosine molecules is counted less than those of proper matches. The highest scoring nucleation point is chosen as the result. The results using degenerated probes were in line with this NucZip theory. A long matching segment created by a wobble position increased hybridization strength beyond the contribution of the additional single match.

At present, the NucZip model is intended for equally long nt segments. Exceptions to the model found in the experimental section of this work can be illustrated by the remarkable difference in hybridization between the short perfectly matching targets of 15–22 nt (15nt_free, 18nt_free and 22nt_free) and the longer 70-mer targets (74%_15nt, 74%_18nt and 74%_22nt) (Figure 3A, B and D). Several groups have analyzed the effect of short so-called dangling ends of 1–5 nt (23–26), and report that they appear to stabilize hybridization of the duplex. Doctycz *et al.* observed that the dangling nt closest to the duplex contributed most to the stabilizing effect (47). The lengths of our long mismatching ends ranged from 55 to 48 nt with a 74%-nt mismatch and no matching trimers. We speculate that the destabilization seen in a duplex with two long mismatching ends, one from the probe and one from the target, compared with a duplex between a long and a short oligonucleotide, with only one long protruding oligonucleotide, is due to shearing stress on the matching nt, or to competition from intra-strand secondary structures at the separate ends.

Although it can predict many aspects of long oligonucleotide hybridization in 3 M TMAC at 45°C, the NucZip concept has to be amended to include other temperatures, oligonucleotide concentrations and buffers in order to be generally useful. The NN model, which was developed with great precision by SantaLucia *et al.* (8,20–22) and is used in the Visual OMP™ 7.0 computer program, is much more sophisticated than our procedure. It was considered out of the scope of this article to evaluate the Visual OMP™ 7.0 program in detail; however, importantly, its hybridization conditions can be adjusted. When Visual OMP™ 7.0 was used to calculate ΔG (Figure 6), the buffer and temperature conditions were set at 3 M TMAC and 45°C. Comparison of the NucZip and Visual OMP™ 7.0 models showed that the distribution of the targets with the long dangling ends (74%_Xnt) was corrected to a certain extent by NucZip; however, neither NucZip nor Visual OMP 7.0 accurately predicted the hybridization behavior of the short perfectly

matching probe–target combinations of the InflA probe with the 15nt_free, 18nt_free and 22nt_free targets.

In conclusion, we have demonstrated that the distribution of mismatches greatly affects probe hybridization. A minimum number of continuous, perfectly matching stretches (nucleation sites) is needed to initiate hybridization. Thus, if the target contains many variations and no long uninterrupted matching segments, use of a dInosine-containing probe, which partially overcomes the obstacles caused by mismatches, will be beneficial. With respect to hybridization prediction algorithms, a simple statement of percentage mismatch, as used in many such algorithms, does not adequately reflect the hybridization properties of a long duplex. The insertion of dInosines in the positions of variation could detect target nucleic acid that a consensus probe would fail to catch. Hence, while high dInosine content in a probe decreases the probe's binding capacity, it also covers mismatches, as required when creating mismatch-tolerant, broad-detection probes against highly variable target nucleic acid sequences such as those seen in RNA viruses. The dInosine probes can be made even more forgiving by using a lower hybridization temperature. Furthermore, the probability of cross hybridization is low because the moieties neighboring the dInosines have a destabilizing effect on hybridization. While the aim of this study was to improve our understanding of variability and the effects of dInosines, other universal bases and wobbles in the design of probes to be used in a 3 M TMAC buffer system, with more exploratory work the results may also be relevant to other hybridization systems and could aid the development of hybridization-based diagnostic tools, including nanotechnological applications such as the volume-amplified magnetic nanobead detection assay (48,49).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Knut and Alice Wallenberg foundation (grant # KAW2007.0130); Academic Hospital, Uppsala, and the Research Council of the Uppland/Västmanland/Örebro/Dalarna region (grant # RFR-86541). Funding for open access charge: Knut and Alice Wallenberg foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Crick, F.H. (1966) Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, **19**, 548–555.
2. Karran, P. and Lindahl, T. (1978) Enzymatic excision of free hypoxanthine from polydeoxynucleotides and DNA containing deoxyinosine monophosphate residues. *J. Biol. Chem.*, **253**, 5877–5879.
3. Karran, P. and Lindahl, T. (1980) Hypoxanthine in deoxyribonucleic acid: generation by heat-induced hydrolysis of adenine residues and release in free form by a deoxyribonucleic acid glycosylase from calf thymus. *Biochemistry*, **19**, 6005–6011.

4. Clutterbuck, D.R., Leroy, A., O'Connell, M.A. and Semple, C.A. (2005) A bioinformatic screen for novel A-I RNA editing sites reveals recoding editing in BC10. *Bioinformatics*, **21**, 2590–2595.
5. Bergstrom, D.E., Zhang, P. and Johnson, W.T. (1997) Comparison of the base pairing properties of a series of nitroazole nucleobase analogs in the oligodeoxyribonucleotide sequence 5'-d(CGCAATTYGCG)-3'. *Nucleic Acids Res.*, **25**, 1935–1942.
6. Loakes, D. (2001) Survey and summary: The applications of universal DNA base analogues. *Nucleic Acids Res.*, **29**, 2437–2447.
7. Martin, F.H., Castro, M.M., Aboul-ela, F. and Tinoco, I. Jr (1985) Base pairing involving deoxyinosine: implications for probe design. *Nucleic Acids Res.*, **13**, 8927–8938.
8. Watkins, N.E. Jr and SantaLucia, J. Jr (2005) Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes. *Nucleic Acids Res.*, **33**, 6258–6267.
9. Loakes, D., Brown, D.M., Linde, S. and Hill, F. (1995) 3-Nitropyrrole and 5-nitroindole as universal bases in primers for DNA sequencing and PCR. *Nucleic Acids Res.*, **23**, 2361–2366.
10. Patil, R.V. and Dekker, E.E. (1990) PCR amplification of an *Escherichia coli* gene using mixed primers containing deoxyinosine at ambiguous positions in degenerate amino acid codons. *Nucleic Acids Res.*, **18**, 3080.
11. Shapiro, J.T., Stannard, B.S. and Felsenfeld, G. (1969) The binding of small cations to deoxyribonucleic acid. Nucleotide specificity. *Biochemistry*, **8**, 3233–3241.
12. Melchior, W.B. Jr and Von Hippel, P.H. (1973) Alteration of the relative stability of dA-dT and dG-dC base pairs in DNA. *Proc. Natl Acad. Sci. USA*, **70**, 298–302.
13. Wood, W.I., Gitschier, J., Lasky, L.A. and Lawn, R.M. (1985) Base composition-independent hybridization in tetramethylammonium chloride: a method for oligonucleotide screening of highly complex gene libraries. *Proc. Natl Acad. Sci. USA*, **82**, 1585–1588.
14. Jacobs, K.A., Rudersdorf, R., Neill, S.D., Dougherty, J.P., Brown, E.L. and Fritsch, E.F. (1988) The thermal stability of oligonucleotide duplexes is sequence independent in tetraalkylammonium salt solutions: application to identifying recombinant DNA clones. *Nucleic Acids Res.*, **16**, 4637–4650.
15. Majlessi, M. and Becker, M.M. (2008) Formation of the double helix: a mutational study. *Nucleic Acids Res.*, **36**, 2981–2989.
16. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
17. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
18. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
19. SantaLucia, J. Jr (2007) Physical principles and visual-OMP software for optimal PCR design. *Methods Mol. Biol.*, **402**, 3–34.
20. Peyret, N., Seneviratne, P.A., Allawi, H.T. and SantaLucia, J. Jr (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. *Biochemistry*, **38**, 3468–3477.
21. Allawi, H.T. and SantaLucia, J. Jr (1998) Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. *Biochemistry*, **37**, 9435–9444.
22. Allawi, H.T. and SantaLucia, J. Jr (1998) Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. *Biochemistry*, **37**, 2170–2179.
23. Bommarito, S., Peyret, N. and SantaLucia, J. Jr (2000) Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res.*, **28**, 1929–1934.
24. Doktycz, M.J., Paner, T.M., Amaratunga, M. and Benight, A.S. (1990) Thermodynamic stability of the 5' dangling-ended DNA hairpins formed from sequences 5'-(XY)2GGATAC(T)4GTATCC-3', where X, Y = A, T, G, C. *Biopolymers*, **30**, 829–845.
25. Senior, M., Jones, R.A. and Breslauer, K.J. (1988) Influence of dangling thymidine residues on the stability and structure of two DNA duplexes. *Biochemistry*, **27**, 3879–3885.
26. Quartin, R.S. and Wetmur, J.G. (1989) Effect of ionic strength on the hybridization of oligodeoxynucleotides with reduced charge due to methylphosphonate linkages to unmodified oligodeoxynucleotides containing the complementary sequence. *Biochemistry*, **28**, 1040–1047.
27. Loakes, D. and Brown, D.M. (1994) 5-Nitroindole as a universal base analogue. *Nucleic Acids Res.*, **22**, 4039–4043.
28. Fadeel, B., Kasemo, B., Malmsten, M. and Stromme, M. (2010) Nanomedicine: reshaping clinical practice. *J. Int. Med.*, **267**, 2–8.
29. He, Z., Wu, L., Li, X., Fields, M.W. and Zhou, J. (2005) Empirical establishment of oligonucleotide probe design criteria. *Appl. Environ. Microbiol.*, **71**, 3753–3760.
30. Kane, M.D., Jatkoe, T.A., Stumpf, C.R., Lu, J., Thomas, J.D. and Madore, S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.
31. He, Z., Wu, L., Fields, M.W. and Zhou, J. (2005) Use of microarrays with different probe sizes for monitoring gene expression. *Appl. Environ. Microbiol.*, **71**, 5154–5162.
32. Deng, Y., He, Z., Van Nostrand, J.D. and Zhou, J. (2008) Design and analysis of mismatch probes for long oligonucleotide microarrays. *BMC Genomics*, **9**, 491.
33. Rennie, C., Noyes, H.A., Kemp, S.J., Hulme, H., Brass, A. and Hoyle, D.C. (2008) Strong position-dependent effects of sequence mismatches on signal ratios measured using long oligonucleotide microarrays. *BMC Genomics*, **9**, 317.
34. Letowski, J., Brousseau, R. and Masson, L. (2004) Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. *J. Microbiol. Methods*, **57**, 269–278.
35. Naiser, T., Ehler, O., Kayser, J., Mai, T., Michel, W. and Ott, A. (2008) Impact of point-mutations on the hybridization affinity of surface-bound DNA/DNA and RNA/DNA oligonucleotide-duplexes: comparison of single base mismatches and base bulges. *BMC Biotechnol.*, **8**, 48.
36. Naiser, T., Kayser, J., Mai, T., Michel, W. and Ott, A. (2008) Position dependent mismatch discrimination on DNA microarrays – experiments and model. *BMC Bioinformatics*, **9**, 509.
37. Hagan, M. and Chakraborty, A. (2004) Hybridization dynamics of surface immobilized DNA. *J. Chem. Phys.*, **120**, 4958–4968.
38. Peterson, A., Heaton, R. and Georgiadis, R. (2000) Kinetic control of hybridization in surface immobilized DNA monolayer films. *J. Am. Chem. Soc.*, **122**, 7837.
39. Peterson, A., Wolf, L. and Georgiadis, R. (2002) Hybridization of mismatched or partially matched DNA at surfaces. *J. Am. Chem. Soc.*, **124**, 14601.
40. Honore, B., Madsen, P. and Leffers, H. (1993) The tetramethylammonium chloride method for screening of cDNA libraries using highly degenerate oligonucleotides obtained by backtranslation of amino-acid sequences. *J. Biochem. Biophys. Methods*, **27**, 39–48.
41. Muradrasoli, S., Mohamed, N., Hornyak, A., Fohlman, J., Olsen, B., Belak, S. and Blomberg, J. (2009) Broadly targeted multiprobe QPCR for detection of coronaviruses: Coronavirus is common among mallard ducks (*Anas platyrhynchos*). *J. Virol. Methods*, **159**, 277–287.
42. Elfaitouri, A., Hammarin, A.L. and Blomberg, J. (2006) Quantitative real-time PCR assay for detection of human polyomavirus infection. *J. Virol. Methods*, **135**, 207–213.
43. Bustin, S.A. (2004) *A-Z of quantitative PCR*. International University Line, La Jolla, Calif.
44. Lutter, L.C. (1979) Precise location of DNase I cutting sites in the nucleosome core determined by high resolution gel electrophoresis. *Nucleic Acids Res.*, **6**, 41–56.
45. SantaLucia, J. Jr, Allawi, H.T. and Seneviratne, P.A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.

46. Hooyberghs,J., Van Hummelen,P. and Carlon,E. (2009) The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters. *Nucleic Acids Res.*, **37**, e53.
47. Doktycz,M.J., Goldstein,R.F., Paner,T.M., Gallo,F.J. and Benight,A.S. (1992) Studies of DNA dumbbells. I. Melting curves of 17 DNA dumbbells with different duplex stem sequences linked by T4 endloops: evaluation of the nearest-neighbor stacking interactions in DNA. *Biopolymers*, **32**, 849–864.
48. Strömberg,M., Göransson,J., Gunnarsson,K., Nilsson,M., Svedlindh,P. and Strømme,M. (2008) Sensitive molecular diagnostics using volume-amplified magnetic nanobeads. *Nano Lett.*, **8**, 816–821.
49. Strömberg,M., Zardán Gómez de la Torre,T., Göransson,J., Gunnarsson,K., Nilsson,M., Strømme,M. and Svedlindh,P. (2008) Microscopic mechanisms influencing the volume amplified magnetic nanobead detection assay. *Biosens. Bioelectron.*, **24**, 696–703.