JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 5, 2008

Pp. 525–534
DOI: 10.1089/cmb.2007.0208

New, Improved, and Practical k-Stem Sequence Similarity Measures for Probe Design

ANTHONY J. MACULA,[1] ALEXANDER SCHLIEP,[2] MORGAN A BISHOP,[1]
and THOMAS E. RENZ[3]

ABSTRACT

We define new measures of sequence similarity for oligonucleotide probe design. These new measures incorporate the nearest neighbor k-stem motifs in their definition, but can be efficiently computed by means of a bit-vector method. They are not as computationally costly as algorithms that predict nearest neighbor hybridization potential. Our new measures for sequence similarity correlate significantly better with nearest neighbor thermodynamic predictions than either BLAST or the standard edit or insertion-deletion defined similarities already in use in many different probe design applications.

Key words: bit-vector, BLAST, edit, hamming, insertion-deletion, nearest neighbor, probe, sequence similarity, target.

1. INTRODUCTION

MANY PAPERS AND SOFTWARE PRODUCTS address the problem of designing relatively short oligonucleotide probes to identify and distinguish oligonucleotide targets by virtue of the targets hybridization signatures with the probes. Frequently, probes are affixed to a surface, for example, microarrays or beads (Cai et al., 2000; Kaderali et al., 2003). There are three general principles that are important in the design of oligonucleotide probes for target identification and discrimination (Nordberg, 2005):

1. Sensitivity: Probes must bind to their intended targets very strongly. Therefore, the hybridization potential of a probe-intended target duplex must be above some sensitivity threshold.
2. Specificity: Probes must not bind to non-targets or other probes. Therefore, the hybridization potential of all probe-non-target and all probe-probe duplexes must be below some sensitivity threshold.
3. Consistency: All probe-intended target duplexes should have similar melting temperatures.

In the computational intensive search for good probes, the notion of *sequence similarity* is important, especially in the preprocessing or probe candidate filter stage that is common to many probe design

[1]Biomathematics Group, SUNY Geneseo, Geneseo, New York.
[2]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany.
[3]IFTC, Rome Research Site, Air Force Research Laboratory, Rome, New York.

applications (Kane et al., 2000; Li et al., 2005). For sensitivity, the Watson-Crick complement of a probe must be quite similar (frequently identical) in its sequence composition to some region of *its and only its* targets. For specificity, the Watson-Crick complement of a probe must be quite dissimilar in its sequence composition to all other probes as well as all regions of all of its unintended targets. Most probe design applications seek a bijection between probes and intended targets. In some cases, there can be multiple probe/intended target duplexes (Schliep et al., 2003).

2. HAMMING, EDIT, AND INSERTION-DELETION SIMILARITY

Throughout this paper, we assume all probes have length m, all targets have length n, and $m \leq n$. Let $x = x_1, \ldots, x_m$ and $y = y_1, \ldots, y_n$ be sequences over a finite alphabet. When x and y are DNA strands, we think of x as a probe and y as a target. For sequence y, we let $y_{i..j}$ denote the substring $y_i, \ldots, y_j$. The *length of the longest common subsequence* between x and y is denoted by $lcs(x, y)$. For $n = m$, the *hamming score*, denoted by $hs(x, y)$, is the number of *corresponding matches* between x and y. The *insertion-deletion score* between x and y is minimum number of insertions or deletions required to change x into y and is denoted by $ids(x, y)$. Similarly, the *edit score* is the minimum number of insertions, deletions or substitutions required to change x into y and is denoted by $eds(x, y)$. Since a single substitution is equivalent to one insertion and one deletion, it follows that $ids(x, y) \geq eds(x, y)$. Note that $ids(x, y) = m + n - 2 \cdot lcs(x, y)$.

Definition 1. *For $m \leq n$, the hamming distance, respectively denoted $h(x, y)$, between x and y is*

$$h(x, y) \equiv \min\{n - hs(x, y_{i..i+n-1}) : 1 \leq i \leq n - m + 1\}.$$

The edit and indel distances, denoted $ed(x, y)$ and $id(x, y)$ between x and y are

$$ed(x, y) \equiv \min\{eds(x, y_{i..j}) : 1 \leq i \leq j \leq n\}$$

$$id(x, y) \equiv \min\{ids(x, y_{i..j}) : 1 \leq i \leq j \leq n\}.$$

These definitions make the notions seem more complicated then they are. $h(x, y)$ is the fewest number corresponding differences between x and some *substring of y of length m*. $ed(x, y)$ and $id(x, y)$ are, respectively, the fewest number of either insertions, deletions, substitutions or just insertions, deletions to be made in x so that it is equal to some *substring* of y of any length not necessarily m.

A table that outlines the functionality of several probe design software applications is given in Nordberg (2005). Many programs listed there use BLAST to determine sequence similarity likely because BLAST is computationally efficient and familiar, but there is a trade off between speed and accuracy. Programs that use BLAST to determine sequence similarity are essentially using $h(x, y)$ (or a translation of it) to measure the similarity between two nucleotide sequences. However, as discussed in Nordberg (2005), there are instances, albeit infrequent, where BLAST does not accurately compute its version of $h(x, y)$. Moreover, as we show in Tables 1 and 2, there may not be a strong correlation between BLAST measures of sequence similarity and nearest neighbor hybridization potential. Not all probe design packages use BLAST. Other probe design packages use a notion of sequence similarity that is based on either $ed(x, y)$ or $id(x, y)$ (Li et al., 2005; Wu et al., 2003).

There are conflicting results about whether or not the aqueous nearest neighbor model (Mathews et al., 2006; SantaLucia, 1998; SantaLucia and Hicks, 2004; Zuker et al., 1999) is an appropriate one for probe design (Li and Stormo, 2001; Zhang et al., 2007; Pozhitkov et al., 2006, 2007; Fish et al., 2007):

> The point of this note is to suggest other measures of sequence similarity that better reflect the nearest neighbor thermodynamic model for hybridization potential of two nucleotide sequences.

While it is possible to use the nearest neighbor thermodynamic computations as a basis to define sequence similarity, they are too computationally costly to be implemented as the only similarity measure used if the amount of sequence data is too large. Our new sequence similarity definitions are based on simple

TABLE 1. THE r^2 VALUES FOR SIMILARITY MEASURES VERSUS PAIRFOLD PREDICTIONS FOR PROBE LENGTH 20

m, n	*Similarity*	$AT =$ *0.15*	$AT =$ *0.25*	$AT =$ *0.35*	*Meiobenthos*	m, n	*Similarity*	$AT =$ *0.15*	$AT =$ *0.25*	$AT =$ *0.35*	*Meiobenthos*
$m = 20$,	$H(x : y)$	*0.059*	*0.004*	*0.000*	*0.005*	$m = 20$,	$H(x : y)$	*0.261*	*0.071*	*0.026*	*0.365*
$n = 20$	$ED(x : y)$	0.302	0.127	0.073	0.248	$n = 30$	$ID(x : y)$	0.280	0.114	0.073	0.532
	$ID(x : y)$	0.261	0.128	0.082	0.295		$ED(x : y)$	0.338	0.136	0.075	0.561
	$ID_2(x : y)$	0.333	0.191	0.129	0.490		$ID_2(x : y)$	0.299	0.153	0.116	0.600
	$ED_2(x : y)$	0.415	0.226	0.142	0.501		$ED_2(x : y)$	0.407	0.202	0.132	0.655
	$ID_5(x : y)$	0.359	0.234	0.239	0.659		$ID_3(x : y)$	0.366	0.240	0.159	0.694
	$ED_5(x : y)$	0.368	0.235	0.241	0.675		$ED_3(x : y)$	**0.433**	0.260	0.176	0.706
	$ID_3(x : y)$	0.419	0.282	0.183	0.626		$ID_5(x : y)$	0.337	0.262	0.222	0.723
	$ID_4(x : y)$	0.433	0.303	0.239	0.686		$ED_5(x : y)$	0.356	0.262	**0.225**	0.750
	$ED_3(x : y)$	**0.475**	0.308	0.206	0.626		$ID_4(x : y)$	0.369	0.280	0.212	0.724
	$ED_4(x : y)$	0.470	**0.320**	**0.246**	**0.708**		$ED_4(x : y)$	0.418	**0.288**	0.216	**0.755**
$m = 20$,	$H(x : y)$	*0.186*	*0.037*	*0.019*	*0.397*	$m = 20$,	$H(x : y)$	*0.157*	*0.028*	*0.017*	*0.387*
$n = 40$	$ID(x : y)$	0.189	0.061	0.043	0.536	$n = 50$	$ID(x : y)$	0.179	0.044	0.032	0.542
	$ID_2(x : y)$	0.225	0.076	0.063	0.596		$ED(x : y)$	0.185	0.054	0.040	0.559
	$ED(x : y)$	0.223	0.080	0.052	0.551		$ID_2(x : y)$	0.194	0.062	0.047	0.579
	$ED_2(x : y)$	0.284	0.120	0.076	0.630		$ED_2(x : y)$	0.237	0.091	0.070	0.618
	$ID_3(x : y)$	0.266	0.123	0.104	0.661		$ID_3(x : y)$	0.220	0.091	0.078	0.636
	$ED_3(x : y)$	**0.314**	0.144	0.109	0.679		$ID_5(x : y)$	0.234	0.111	0.129	0.648
	$ID_5(x : y)$	0.252	0.145	0.154	0.675		$ED_3(x : y)$	0.268	0.113	0.095	0.656
	$ED_5(x : y)$	0.266	0.147	**0.157**	0.710		$ED_5(x : y)$	0.244	0.113	**0.130**	0.683
	$ID_4(x : y)$	0.270	0.160	0.140	0.677		$ID_4(x : y)$	0.240	0.118	0.114	0.647
	$ED_4(x : y)$	0.307	**0.168**	0.145	0.715		$ED_4(x : y)$	**0.268**	**0.125**	0.120	**0.683**

The probe and target lengths are m and n, respectively. The columns, A, $T = p$ with $0 \leq p \leq 1$ indicate that strands were randomly generation where $\text{prob}(A) = \text{prob}(T) = p$ and $\text{prob}(C) = \text{prob}(G) = 1 - p$. The column "meiobenthos" indicates oligos selected from meiobenthos (i.e., small benthic invertebrates that live in both marine and fresh water environments) genomic sequences. The r^2 values are listed in increasing order for the A, $T = 0.25$ column. The maximum and minimum r^2 values in each column are in bold and italics, respectively. It is important to observe that in all cases even $ED_2(x : y)$ is significantly better than either $H(x : y)$, $ED(x : y)$, or $ID(x : y)$ currently in use.

TABLE 2. THE r^2 VALUES FOR SIMILARITY MEASURES VERSUS PAIRFOLD PREDICTIONS FOR PROBE LENGTH 30

m, n	*Similarity*	$AT =$ *0.15*	$AT =$ *0.25*	$AT =$ *0.35*	*Meiobenthos*	m, n	*Similarity*	$AT =$ *0.15*	$AT =$ *0.25*	$AT =$ *0.35*	*Meiobenthos*
$m = 30$,	$H(x : y)$	*0.065*	*0.009*	*0.000*	*0.006*	$m = 30$,	$H(x : y)$	*0.208*	*0.047*	*0.008*	*0.305*
$n = 30$	$ID_2(x : y)$	0.288	0.103	0.095	0.511	$n = 40$	$ID(x : y)$	0.234	0.102	0.057	0.472
	$ID(x : y)$	0.251	0.106	0.066	0.395		$ID_2(x : y)$	0.250	0.105	0.068	0.519
	$ED(x : y)$	0.303	0.124	0.063	0.441		$ED(x : y)$	0.315	0.136	0.071	0.599
	$ID_3(x : y)$	0.316	0.168	0.127	0.575		$ID_3(x : y)$	0.286	0.149	0.099	0.579
	$ED_2(x : y)$	0.405	0.180	0.117	0.610		$ID_5(x : y)$	0.289	0.189	0.184	0.593
	$ID_5(x : y)$	0.308	0.203	0.209	0.601		$ID_4(x : y)$	0.299	0.190	0.147	0.596
	$ED_5(x : y)$	0.347	0.210	**0.218**	0.682		$ED_2(x : y)$	0.377	0.200	0.112	0.662
	$ID_4(x : y)$	0.326	0.212	0.181	0.606		$ED_5(x : y)$	0.329	0.206	**0.195**	0.690
	$ED_3(x : y)$	**0.437**	0.228	0.165	0.674		$ED_3(x : y)$	**0.388**	0.237	0.147	0.688
	$ED_4(x : y)$	0.410	**0.250**	0.205	**0.703**		$ED_4(x : y)$	0.376	**0.241**	0.175	**0.710**
$m = 30$,	$H(x : y)$	*0.162*	*0.036*	*0.018*	*0.470*	$m = 30$,	$H(x : y)$	0.165	*0.027*	*0.012*	*0.381*
$n = 50$	$ID_2(x : y)$	0.199	0.069	0.057	0.530	$n = 60$	$ID(x : y)$	*0.163*	0.051	0.024	0.507
	$ID(x : y)$	0.164	0.076	0.045	0.478		$ID_2(x : y)$	0.177	0.054	0.032	0.545
	$ED(x : y)$	0.221	0.107	0.067	0.576		$ED(x : y)$	0.204	0.076	0.043	0.562
	$ID_3(x : y)$	0.229	0.110	0.084	0.571		$ID_3(x : y)$	0.205	0.077	0.043	0.587
	$ED_2(x : y)$	0.276	0.134	0.098	0.640		$ED_2(x : y)$	0.247	0.101	0.063	0.633
	$ID_4(x : y)$	0.238	0.142	0.121	0.583		$ID_4(x : y)$	0.198	0.108	0.086	0.587
	$ID_5(x : y)$	0.227	0.142	0.152	0.576		$ED_3(x : y)$	0.276	0.113	0.073	0.670
	$ED_5(x : y)$	0.264	0.151	**0.158**	0.666		$ID_5(x : y)$	0.193	0.119	0.103	0.579
	$ED_3(x : y)$	**0.303**	0.161	0.118	0.671		$ED_5(x : y)$	0.226	0.121	**0.109**	0.674
	$ED_4(x : y)$	0.299	**0.169**	0.146	**0.683**		$ED_4(x : y)$	**0.261**	**0.124**	0.094	**0.690**

See Table 1 for legend.

generalizations of the edit and insertion-deletion distances, and can be implemented in a computationally efficient way by means of the bit-vector approach to computing a dynamic programming matrix (Allison and Dix, 1986; Myers, 1999; Crochemore et al., 2001; Hyyrö et al., 2005).

3. SIMILARITY MEASURES BASED ON NUCLEOTIDE ALIGNMENT

Single strands of DNA are represented by *(A, C, G, T)*–quaternary sequences that are oriented, either $5' \to 3'$ or $3' \to 5'$. In this paper, *single-stranded* DNA molecules without an indicated direction are assumed to be in the $5' \to 3'$ direction. The *reverse-complement* of a DNA strand is defined by first reversing the order of the letters and then substituting each letter with its complement, A for T, C for G, and vice-versa. For example, the reverse complement of *AACGTG* is *CACGTT*. Henceforth, *complement* means reverse-complement. For strand x, let $\overline{x}$ denote its complement. A *Watson-Crick duplex* is the joining of complement sequences in opposite orientations so that every base of one strand is paired with its complementary base on the other strand in the double helix structure, i.e., x and $\overline{x}$ are "perfectly similar." However, when two, not necessarily complementary, oppositely directed DNA strands are "sufficiently similar," they too are capable of coalescing into a double-stranded DNA duplex. The process of forming DNA duplexes from single strands is referred to as *DNA hybridization. Crosshybridization* is when two oppositely directed and non-complementary DNA strands form a duplex. Crosshybridization doesn't always occur, but there is a potential for it to happen. In general, crosshybridization is undesirable as it usually leads to experimental error. In general, a collection of probes has sufficient specificity if no crosshybridization takes place. However, there are situations where probes for intended targets are not perfectly complementary to intended binding regions (Karaman et al., 2005).

Definition 2. *Given two DNA strands x and y, we let $x : y$ denote the DNA duplex formed between x and y. It is implicitly assumed that x and y are oppositely oriented in $x : y$ with the first strand x* always *assumed to be in the $5' \to 3'$ direction and the second y* always *assumed to be in the $3' \to 5'$ direction.*

Let $x = x_1, \ldots, x_m$ and $y = y_1, \ldots, y_n$ be DNA sequences. For a base y_j, let $\tilde{y}_j$ be its complement base. Then $\overline{y} = \tilde{y}_m, \ldots, \tilde{y}_1$. A natural simplification for formulating binding specificity is to base it upon the maximum number of base pair bonds between complementary letter pairs in the $x : y$ duplex. An upper bound on the maximum number of inter-strand base pair bonds that can form in the $x : y$ duplex (without pseudoknots) is the maximum length of a common subsequence of x and $\overline{y}$. In short, two single stranded DNA sequences x and y of length m and n can form d inter-strand base pairs bonds in a duplex only if $lcs(x, \overline{y}) \geq d$. See Example 2 below. This, coupled with Definition 1 leads to three known and applied definitions of similarity scores for a duplex.

Definition 3. *For $m \leq n$, and let x and y be DNA sequences. Then the hamming, edit and insertion-deletion scores for the $x : y$ duplex, denoted by $Hs(x : y), EDs(x : y)$ and $IDs(x : y)$, are defined as*

$$Hs(x : y) \equiv hs(x, \overline{y})$$

$$EDs(x : y) \equiv eds(x, \overline{y})$$

$$IDs(x : y) \equiv ids(x, \overline{y}).$$

The hamming, edit, and insertion-deletion similarities for the $x : y$ duplex, denoted by $H(x : y)$, $ED(x : y)$, and $ID(x : y)$, are

$$H(x : y) \equiv h(x, \overline{y})$$

$$ED(x : y) \equiv ed(x, \overline{y})$$

$$ID(x : y) \equiv id(x, \overline{y}).$$

	$\bar{y}$	G	C	G	A	G	T	T	T	A	T	G	C	C	A
x	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1
C	2	2	1	1	2	2	2	2	2	2	2	2	1	0	1
C	3	3	2	2	2	3	3	3	3	3	3	3	2	1	1
A	4	4	3	3	2	3	4	4	4	3	4	4	3	2	1
G	5	4	4	3	3	2	3	4	5	4	4	4	4	3	2
T	6	5	5	4	4	3	2	3	4	5	4	5	5	4	3
T	7	6	6	5	5	4	3	2	3	4	5	5	6	5	4
T	8	7	7	6	6	5	4	3	2	3	4	5	6	6	5
T	9	8	8	7	7	6	5	4	3	3	3	4	5	6	6
A	10	9	9	8	7	7	6	5	4	3	4	4	5	6	6
C	11	10	9	9	8	8	7	6	5	4	4	5	4	5	6
C	12	11	10	10	9	9	8	7	6	5	5	5	5	4	5
C	13	12	11	11	10	10	9	8	7	6	6	6	5	5	5
C	14	13	12	12	11	11	10	9	8	7	7	7	6	5	6

(a)

	$\bar{y}$	G	C	G	A	G	T	T	T	A	T	G	C	C	A
x	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1
C	2	2	1	2	2	2	2	2	2	2	2	2	1	0	1
C	3	3	2	3	3	3	3	3	3	3	3	3	2	1	2
A	4	4	3	4	3	4	4	4	4	3	4	4	3	2	1
G	5	4	4	3	4	3	4	5	5	4	5	4	4	3	2
T	6	5	5	4	5	4	3	4	5	5	4	5	5	4	3
T	7	6	6	5	6	5	4	3	4	5	5	6	6	5	4
T	8	7	7	6	7	6	5	4	3	4	5	6	7	6	5
T	9	8	8	7	8	7	6	5	4	5	4	5	6	7	6
A	10	9	9	8	7	8	7	6	5	4	5	6	7	8	7
C	11	10	9	9	8	9	8	7	6	5	6	7	6	7	8
C	12	11	10	10	9	10	9	8	7	6	7	8	7	6	7
C	13	12	11	11	10	11	10	9	8	7	8	9	8	7	8
C	14	13	12	12	11	12	11	10	9	8	9	10	9	8	9

(b)

FIG. 1. An example of edit dynamic programming matrices *EDM* and *IDM*.

Note that it follows that $IDs(x : y) = m + n - 2 \cdot lcs(x, \overline{y})$. To compute $ed(x, y)$ and $id(x, y)$ it is not necessary to compute $eds(x, y_{i..j})$ and $ids(x, y_{i..j})$ for each $y_{i..j}$ with $1 \leq i \leq j \leq n$. The value of $ed(x, y)$ and $id(x, y)$ can be found by populating a recursively defined $(m + 1) \times (n + 1)$ dynamic programming matrix. Let $EDM(x, y)$ and $IDM(x, y)$ be $(m + 1) \times (n + 1)$ non-negative integer matrices recursively defined by

$$EDM_{i,0} = IDM_{i,0} = i \text{ for } 0 \leq i \leq m$$

$$EDM_{0,j} = IDM_{0,j} = 0 \text{ for } 0 \leq j \leq n$$

$$EDM_{i,j} = \begin{cases} EDM_{i-1,j-1} \text{ if } x_i = y_j \\ 1 + \min\{EDM_{i-1,j}, EDM_{i-1,j-1}, EDM_{i,j-1}\} \text{ otherwise} \end{cases}$$

$$IDM_{i,j} = \begin{cases} IDM_{i-1,j-1} \text{ if } x_i = y_j \\ 1 + \min\{IDM_{i-1,j}, IDM_{i,j-1}\} \text{ otherwise.} \end{cases}$$

Then

$$ed(x, y) = \min\{EDM_{m,j} : 0 \leq j \leq m\}$$

$$id(x, y) = \min\{IDM_{m,j} : 0 \leq j \leq m\}.$$

See Gusfield (1997), Myers (1999), and Hyyrö et al. (2005).

Example 1. *Let* $x = CCCAGTTTTACCCC$ *and* $y = TGGCATAAACTCGC$. *Then* $\overline{y} = GCGAGTTTATGCCA$. *So to compute* $ED(x : y)$ *and* $ID(x : y)$ *the matrices* $EDM(x, \overline{y})$ *and* $IDM(x, \overline{y})$ *are given in Figures 1a and 1b, respectively. Since in Figures 1a and 1b, the minimum entries along the bottom rows are 5 and 8, respectively, we have that* $ED(x : y) = 5$ *and* $ID(x : y) = 8$.

4. SIMILARITY MEASURES BASED ON STACKED k-STEMS ALIGNMENTS

Definition 4. *Suppose* $1 \leqslant i_r, j_r \leqslant m$. *A* secondary structure *of the DNA duplex* $x : y$ *is a sequence of pairs of complementary bases* $\rho = (x_{i_r}, y_{n+1-j_r})$ *where* $x_{i_r} = \tilde{y}_{n+1-j_r}$ *and* (x_{i_r}) *and* (y_{n+1-j_r}) *are increasing and decreasing subsequences of* x *and* y, *respectively. For* $k \geq 2$, *given a secondary structure*

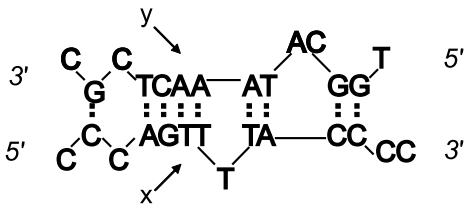

FIG. 2. An example of a secondary structure in a DNA duplex.

$\rho = (x_{i_r}, y_{n+1-j_r})$, *a* stacked k-stem *in a duplex is a k-tuple of consecutively aligned complementary bases,* $x_{i_r} = \tilde{y}_{n+1-j_r}, x_{i_{r+1}} = \tilde{y}_{n+1-j_{r+1}}, \ldots, x_{i_{r+k-1}} = \tilde{y}_{n+1-j_{r+k-1}}$ *in* ρ *where* $i_{r+l-1} = i_r + l - 1$ *and* $j_{r+l-1} = j_r + l - 1$ *for all* l *with* $1 \leq l \leq k - 1$. *Stacked 2-stems and 3-stems are also referred to as stacked pairs and stacked triples, respectively.*

Clearly, the $x : y$ duplex can have many secondary structures; thus, stacked k-stems must be defined relative to a given secondary structure.

Example 2. *The secondary structure in Figure 2 has nine complementary base pairing. This is the most possible since, as indicated in Figure 3, the* $lcs(x, \overline{y}) = 9$. *Similar to Example 1, the value of* $lcs(x, y)$ *can also be found by populating a recursively defined* $(m + 1) \times (n + 1)$ *dynamic programming matrix. Let* $LCSM(x, y)$ *be* $(m + 1) \times (n + 1)$ *non-negative integer matrix recursively defined by*

$$LCSM_{i,0} = LCSM_{j,0} = 0 \text{ for } 0 \leq i \leq m \text{ and } 1 \leq j \leq n$$

$$LCSM_{i,j} = \begin{array}{l} LCSM_{i-1,j-1} + 1 \text{ if } x_i = y_j \\ \max\{LCSM_{i-1,j}, LCSM_{i-1,j-1}, LCSM_{i,j-1}\} \text{ otherwise.} \end{array}$$

Then

$$lcs(x, y) = LCSM_{m,n}.$$

$LCSM(x, \overline{y})$ *for* x *and* y *in Figure 2 is given in Figure 3.*

	$\bar{y}$	G	C	G	A	G	T	T	T	A	T	G	C	C	A
x	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
C	0	0	1	1	1	1	1	1	1	1	1	1	2	2	2
C	0	0	1	1	1	1	1	1	1	1	1	1	2	3	3
A	0	0	1	1	2	2	2	2	2	2	2	2	2	3	4
G	0	1	1	2	2	3	3	3	3	3	3	3	3	3	4
T	0	1	1	2	2	3	4	4	4	4	4	4	4	4	4
T	0	1	1	2	2	3	4	5	5	5	5	5	5	5	5
T	0	1	1	2	2	3	4	5	6	6	6	6	6	6	6
T	0	1	1	2	2	3	4	5	6	6	7	7	7	7	7
A	0	1	1	2	3	3	4	5	6	7	7	7	7	7	8
C	0	1	2	2	3	3	4	5	6	7	7	7	8	8	8
C	0	1	2	2	3	3	4	5	6	7	7	7	8	9	9
C	0	1	2	2	3	3	4	5	6	7	7	7	8	9	9
C	0	1	2	2	3	3	4	5	6	7	7	7	8	9	9
C	0	1	2	2	3	3	4	5	6	7	7	7	8	9	9

FIG. 3. An example of edit dynamic programming matrix *LCSM*.

Example 3. *The secondary structure given in Figure 2 has five stacked pairs,*

$$A_4G_5/T_{11}C_{10}, \quad G_5T_6/C_{10}A_9, \quad T_6T_7/A_9A_8, \quad T_9A_{10}/A_7T_6, \quad C_{11}C_{12}/G_3G_2,$$

where the subscripts indicate the position of the bases in the $5' \to 3'$ *direction in the* x *and* y *strands respectively. The two stacked triples are*

$$A_4G_5T_6/T_{11}C_{10}A_9, \quad G_5T_6T_7/C_{10}A_9A_8$$

and the single stacked 4-stem is

$$A_4G_5T_6T_7/T_{11}C_{10}A_9A_8.$$

If the hybridization potential were solely dependent on the number of base pair bonds, then using $ED(x : y)$ or $ID(x : y)$ as a measure of specificity would make good sense. However, while the hybridization potential of DNA duplexes depends, in part, on the number of base pair bonds, the state of the art model of DNA duplex thermodynamics is the nearest neighbor model where stacked k-stems play a special role. Briefly, local thermodynamic functions ΔH, ΔS, which are essentially independent of temperature T, are experimentally found for stacked pairs, stacked k-stems, and other secondary structure motifs and are then used, in an additive fashion, to predict global thermodynamic values for duplexes (Mathews et al., 2006; SantaLucia, 1998; SantaLucia and Hicks, 2004; Zuker et al., 1999). The main point is that, in regard to hybridization potential, the alignment of stacked pairs and k-stems is a more important consideration than the alignment of the base pairs.

Thus, from the point of view of hybridization potential, it seems reasonable to think of $x = x_1, \ldots, x_m$ and $y = y_1, \ldots, y_n$ not as sequences of bases, but as sequences of consecutive k-strings that could then be aligned.

Definition 5. *Identify A, C, G, and T with* 0, 1, 2, *and* 3, *respectively. Then for a DNA sequence* $x = x_1, \ldots, x_m$ *let* $\langle x_{i..i+k-1} \rangle$ *be the unique number* $\{0, 1, \ldots, 4^k - 1\}$ *whose q-ary decimal representation is* $x_{i..i+k-1}$. *Let* $x^{(k)}$ *be defined as* $x^{(k)} \equiv (\langle x_{i..i+k-1} \rangle)_{i=1}^{m-k+1}$.

Example 4. *For* $x = 2, 3, 3, 0, 3, 0, 2, 2, 1, 1, 2, 0, 2$, *then*

$$x^{(2)} = 11, 15, 12, 3, 12, 2, 10, 9, 5, 6, 8, 2$$

$$x^{(3)} = 47, 60, 51, 12, 50, 10, 41, 37, 22, 24, 34.$$

By making an analogy to the result that the maximum number of inter-strand base pair bonds that can form in the $x : y$ duplex is $lcs(x, \overline{y})$, one could conjecture that maximum number of stacked k-stems in the $x : y$ duplex is $lcs(x^{(k)}, \overline{y}^{(k)})$. As shown in D'yachkov et al. (2005a, 2005b, 2006), this conjecture is false, and a counter-example is given in Example 5. However, D'yachkov et al. (2006) does show that $lcs(x^{(k)}, \overline{y}^{(k)})$ is an upper bound maximum number of stacked k-stems in the $x : y$ duplex. Specifically, D'yachkov et al. (2006) shows that the *longest common k-gap block isomorphic subsequence* between $x^{(k)}$ and $\overline{y}^{(k)}$ is the maximum number of inter-strand stacked k-stems in the $x : y$ duplex (without pseudoknots). A discussion of k-gap block isomorphic subsequences is beyond the scope of this note.

Example 5. *Consider* $x = AGGAC$ *and* $y = TCTCA$ *Thus,* $\overline{y} = AGAGT$. *Then* $x^{(2)} = (AG, GG, GA, AC) = (2, 10, 8, 1)$ *and* $\overline{y}^{(2)} = (AG, GA, AG, GT) = (2, 8, 2, 11)$. *An example of a longest common subsequence between* $x^{(2)}$ *and* $\overline{y}^{(2)}$ *is* $AG, GA = 2, 8$. *However, the alignment of* $x^{(2)}$ *and* $\overline{y}^{(2)}$ *that matches this common subsequence does not correspond to any alignment of the bases in* x *and* y *that gives two stacked pairs in the* $x : y$ *duplex (Fig. 4).*

However, even though an alignment of $x^{(k)}$ and $\overline{y}^{(k)}$ may not correspond to any actual secondary structure of the $x : y$ duplex, it does capture aspects of secondary structures of the $x : y$ duplex by giving a fairly tight upper bound on the number of k-stems that are possible in an actual secondary structure of the $x : y$ duplex. Hence, we use "alignments" of $x^{(k)}$ and $\overline{y}^{(k)}$ to define the following new measures for hybridization

$\overline{y}^{(2)}$ = AG * GA * AG GT $\quad$ $y^{(2)}$ = T**C** * **C**T * TC CA $\quad$ y = T**C**TCA
$x^{(2)}$ = AG GG GA AC $\quad$ $x^{(2)}$ = A**G** GG **G**A AC $\quad$ x = A**GG**AC

FIG. 4. Two-stem *lcs* without an associated secondary structure. The standard longest common subsequence on 2-stems does not correspond to an actual secondary structure. Two Gs can't both bind to the same C.

specificity. Although it is possible to define analogs of the the edit and insertion-deletion similarities that capture the required "k-gap block isomorphic" properties discussed in D'yachkov et al. (2006), and that would better reflect actual secondary structures, the complexity of the known algorithms that make these "k-gap" computations are too costly. Since practical bit vector algorithms exist to compute the standard edit and insertion-deletion dynamic programming matrices, we make the following definitions.

Definition 6. *For $m \leq n$, and let x and y be DNA sequences. The k-stem edit and insertion-deletion similarities for the $x : y$ duplex, denoted by $ED_k(x : y)$ and $ID_k(x : y)$, are*

$$ED_k(x : y) \equiv ed(x^{(k)}, \overline{y}^{(k)})$$

$$ID_k(x : y) \equiv ed(x^{(k)}, \overline{y}^{(k)}).$$

Example 6. *Let $x = CCCAGTTTACCCC$, $y = TGGCATAAACTCGC$, and $\overline{y} = GCGAGTTTATGCCA$ be as in Example 1. Then $EDM(x^{(2)}, \overline{y}^{(2)})$, $IDM(x^{(2)}, \overline{y}^{(2)})$, and $LCSM(x^{(2)}, \overline{y}^{(2)})$ are given in Figures 5a–5c, respectively. From these figures we have $ED_2(x : y) = 7$, $ID_2(x : y) = 8$ and $LCS(x^{(2)}, \overline{y}^{(2)}) = 6$.*

	ȳ	GC	CG	GA	AG	GT	TT	TT	TA	AT	TG	GC	CC	CA
x	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC	1	1	1	1	1	1	1	1	1	1	1	1	0	1
CC	2	2	2	2	2	2	2	2	2	2	2	2	1	1
CA	3	3	3	3	3	3	3	3	3	3	3	3	2	1
AG	4	4	4	4	3	4	4	4	4	4	4	4	3	2
GT	5	5	5	5	4	3	4	5	5	5	5	5	4	3
TT	6	6	6	6	5	4	3	4	5	6	6	6	5	4
TT	7	7	7	7	6	5	4	3	4	5	6	7	6	5
TT	8	8	8	8	7	6	5	4	4	5	6	7	7	6
TA	9	9	9	9	8	7	6	5	4	5	6	7	8	7
AC	10	10	10	10	9	8	7	6	5	5	6	7	8	8
CC	11	11	11	11	10	9	8	7	6	6	6	7	7	8
CC	12	12	12	12	11	10	9	8	7	7	7	7	7	8
CC	13	13	13	13	12	11	10	9	8	8	8	8	7	8

(a)

	ȳ	GC	CG	GA	AG	GT	TT	TT	TA	AT	TG	GC	CC	CA
x	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC	1	1	1	1	1	1	1	1	1	1	1	1	0	1
CC	2	2	2	2	2	2	2	2	2	2	2	2	1	2
CA	3	3	3	3	3	3	3	3	3	3	3	3	2	1
AG	4	4	4	4	3	4	4	4	4	4	4	4	3	2
GT	5	5	5	5	4	3	4	5	5	5	5	5	4	3
TT	6	6	6	6	5	4	3	4	5	6	6	6	5	4
TT	7	7	7	7	6	5	4	3	4	5	6	7	6	5
TT	8	8	8	8	7	6	5	4	5	6	7	8	7	6
TA	9	9	9	9	8	7	6	5	4	5	6	7	8	7
AC	10	10	10	10	9	8	7	6	5	6	7	8	9	8
CC	11	11	11	11	10	9	8	7	6	7	8	9	8	9
CC	12	12	12	12	11	10	9	8	7	8	9	10	9	10
CC	13	13	13	13	12	11	10	9	8	9	10	11	10	11

(b)

	ȳ	GC	CG	GA	AG	GT	TT	TT	TA	AT	TG	GC	CC	CA
x	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC	0	0	0	0	0	0	0	0	0	0	0	0	1	1
CC	0	0	0	0	0	0	0	0	0	0	0	0	1	1
CA	0	0	0	0	0	0	0	0	0	0	0	0	1	2
AG	0	0	0	0	1	1	1	1	1	1	1	1	1	2
GT	0	0	0	0	1	2	2	2	2	2	2	2	2	2
TT	0	0	0	0	1	2	3	3	3	3	3	3	3	3
TT	0	0	0	0	1	2	3	4	4	4	4	4	4	4
TT	0	0	0	0	1	2	3	4	4	4	4	4	4	4
TA	0	0	0	0	1	2	3	4	5	5	5	5	5	5
AC	0	0	0	0	1	2	3	4	5	5	5	5	5	5
CC	0	0	0	0	1	2	3	4	5	5	5	5	6	6
CC	0	0	0	0	1	2	3	4	5	5	5	5	6	6
CC	0	0	0	0	1	2	3	4	5	5	5	5	6	6

(c)

FIG. 5. Examples of $EDM(x^{(2)}, y^{(2)})$, $IDM(x^{(2)}, y^{(2)})$, and $LCSM(x^{(2)}, y^{(2)})$.

5. DISCUSSION

Using simple linear regression to nearest neighbor predictions, we compared the measures $ED_k(x : y)$, $ID_k(x : y)$, and $H(x : y)$ of similarity, the latter of which essentially reflects the performance of BLAST. To do this, we used both randomly generated DNA sequences and randomly selected genomic sequences. For several combinations of probe and target lengths m and n, 5000 pairs of strands were selected. The nearest neighbor software package PAIRFOLD (Andronescu et al., 2003) was used to predict nearest neighbor hybridization potential of the selected $x : y$ duplex, denoted by $PFE(x : y)$, which is measured in terms of the free energy of duplex formation at 37°C. Because PAIRFOLD considers the formation of intra-strand loops, only duplexes where the most stable secondary structure had no internal loops were used. This seems to be a reasonable assumption to make when designing probes that will not crosshybridize because the intra-strand interactions would tend to decrease the availability of crosshybridizing sites and thus make crosshybridization less likely. Moreover, since the probes for intended targets are almost always completely or nearly complementary to binding regions, one would expected intra-strand interactions to not be competitive with complete inter-strand complementarity.

Example 7. *Tables 1 and 2 indicate the* r^2 *values found for* $ED_k(x : y)$, $ID_k(x : y)$, *and* $H(x : y)$ *measures versus PAIRFOLD* $PF(x : y)$ *measurements. In all randomly generated cases, the Hamming similarity shows no correlation with the nearest neighbor predictions. In all cases, the k-stem similarity measures where* $k \geq 2$ *have significantly higher correlations that either of the simple single base measures* $ED(x : y)$, $ID(x : y)$.

6. CONCLUSION

We have shown that the new k-stem $k \geq 2$ measures for hybridization specificity correlates significantly better with nearest neighbor thermodynamic predictions of hybridization potential than either the hamming specificity found in BLAST or the standard edit or insertion-deletion similarities that are in current use. This is especially true for $ED_k(x : y)$. These new edit distance measures can be implemented by Myers bit-vector method so they are not too costly computational to use. In comparison to both a randomly generated and randomly selected genomic oligos, the $ED_3(x : y)$ or $ED_4(x : y)$ seemed always to be best.

ACKNOWLEDGMENTS

We would like to thank Dan Tulpan and the BETA Lab at the University of British Columbia for providing us with PAIRFOLD software and many helpful discussions. This work has been supported by AFOSR FA8750-07-C-0089.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Allison, L., and Dix, T.L. 1986. A bit-string longest common subsequence algorithm. *Inform. Process. Lett.* 23, 305–310.

Andronescu, M., Aguirre-Hernandez, R., Condon, A., et al. 2003. RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.* 31, 3416–3422.

Cai, H., White, P., Torney, D., et al. 2000. Flow cytometry-based minisequencing: a new platform for high throughput single nucleotide polymorphism scoring. *Genomics* 66, 135–143.

Crochemore, M., Iliopoulos, C.S., Pinzon, Y.J., et al. 2001. A fast and practical bit-vector algorithm for the longest common subsequence problem. *Inform. Process. Lett.* 80, 279–285.

D'yachkov, A., Macula, A., Pogozelski, W., et al. 2005a. A weighted insertion deletion stacked pair thermodynamic metric for DNA codes. *Lect. Notes Comput. Sci.* 3384, 90–103.

Dyachkov, A., Macula, A., Pogozelski, W., et al. 2006. New t-gap insertion-deletion like metrics for DNA hybridization thermodynamic modeling. *J. Comput. Biol.* 13, 866–881.

D'yachkov, A., Vilenkin, P., Ismagilov, I., et al. 2005b. On DNA codes. *Problems Inform. Transm.* 41, 349–367.

Fish, D.J., Horne, M.T., and Brewood, G.P. 2007. DNA multiplex hybridization on microarrays and thermodynamic stability in solution: a direct comparison. *Nucleic Acids Res.* 35, 7197–7208.

Gusfield, D. 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge.

Hyyrö, H., Pinzon, Y., and Shinohara, A. 2005. New bit-parallel indel-distance algorithm. *Lect. Notes Comput. Sci.* 3503, 380–390.

Kaderali, L., Deshpande, A., Nolan, J., et al. 2003. Primer-design for multiplexed genotyping. *Nucleic Acids Res.* 31, 1796–1802.

Kane, M.D., Jatkoe, T.A., Stumpf, C.R., et al., 2000. Assessment of the specificity and sensitivity of oligonucleotide (50mer) microarrays. *Nucleic Acid Res.* 28, 4552–4557.

Karaman, M.W., Groshen, S., Lee, C., et al. 2005. Comparisons of substitution, insertion and deletion probes for resequencing and mutational analysis using oligonucleotide microarrays. *Nucleic Acids Res.* 33, e33.

Li, F., and Stormo, G. 2001. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* 17, 1067–1076.

Li, X., He, Z., and Zhou, J. 2005. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. *Nucleic Acids Res.* 33, 6114–6123.

Mathews, D., Zuker, M., and Turner, D. 2006. RNAstructure 4.2. Available at *http://rna.chem.rochester.edu*. Accessed May 10, 2008.

Myers, E.W. 1999. A fast bit-vector algorithm for approximate stringmatching based on dynamic programming. *J. ACM* 46, 539–553.

Nordberg, E.K. 2005. YODA: selecting signature oligonucleotides. *Bioinformatics* 21, 1365–1370.

Pozhitkov, A., Noble, P.A., Domazet-Los, T., et al. 2006. Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. *Nucleic Acids Res.* 34, e66.

Pozhitkov, A.E., Tautz, D., and Noble, P.A. 2007. Oligonucleotide microarrays: widely applied: poorly understood. *Brie. Funct. Genom. Proteom.* 6, 141–148.

SantaLucia, J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, *Proc. Natl. Acad. Sci. USA* 95, 1460–1465.

SantaLucia, J., and Hicks, D. 2004. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* 33, 415–440.

Schliep, A., Torney, D.C., and Rahmann, S. 2003. Group testing with DNA chips: generating designs and decoding experiments. *Proc. IEEE Comput. Soc. Conf. Bioinform.* 84–91.

Wu, C.T., Liao, C.Y., and Su, H.J. 2003. IMPORT—integrated massive probe's optimal recognition tools. *Genome Inform.* 14, 478–479.

Zhang, L., Wu, C., Carta, R., et al. 2007. Free energy of DNA duplex formation on short oligonucleotide microarrays. *Nucleic Acids Res.* 35, e18.

Zuker, M., Mathews, D., and Turner, D. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: Barciszewski, J., and Clark, B.F.C., eds., *RNA Biochemistry and Biotechnology*. Kluwer Academic Publishers, Amsterdam, pgs. 11–45.

Address reprint requests to:
Dr. Anthony J. Macula
Biomathematics Group
SUNY Geneseo
Geneseo, NY 14454

E-mail: macula@geneseo.edu