BIOINFORMATICS

Vol. 18 no. 10 2002
Pages 1340–1349

Selecting signature oligonucleotides to identify organisms using DNA arrays

Lars Kaderali [1,*] and Alexander Schliep [2]

[1] Center for Applied Computer Sciences Cologne (ZAIK), University of Cologne, Weyertal 80, 50931 Köln, Germany and [2] Department of Computational Molecular Biology, MPI for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Received on December 5, 2001; revised on February 4, 2002; accepted on March 27, 2002

ABSTRACT

Motivation: DNA arrays are a very useful tool to quickly identify biological agents present in some given sample, e.g. to identify viruses causing disease, for quality control in the food industry, or to determine bacteria contaminating drinking water. The selection of specific oligos to attach to the array surface is a relevant problem in the experiment design process. Given a set S of genomic sequences (the target sequences), the task is to find at least one oligonucleotide, called probe, for each sequence in S. This probe will be attached to the array surface, and must be chosen in a way that it will not hybridize to any other sequence but the intended target. Furthermore, all probes on the array must hybridize to their intended targets under the same reaction conditions, most importantly at the temperature T at which the experiment is conducted.

Results: We present an efficient algorithm for the probe design problem. Melting temperatures are calculated for all possible probe–target interactions using an extended nearest-neighbor model, allowing for both non-Watson–Crick base-pairing and unpaired bases within a duplex. To compute temperatures efficiently, a combination of suffix trees and dynamic programming based alignment algorithms is introduced. Additional filtering steps during preprocessing increase the speed of the computation.

The practicability of the algorithms is demonstrated by two case studies: The identification of HIV-1 subtypes, and of 28S rDNA sequences from >400 organisms.

Availability: The software is available on request.

Contact: kaderali@zpr.uni-koeln.de

Supplementary information: http://www.zaik.uni-koeln.de/bioinformatik/arraydesign.html

INTRODUCTION

Efficient diagnostic tests to probe genomic information are of great interest for a wide range of applications, for example in medicine or biology. DNA arrays can probe a large number of targets simultaneously, thus reducing time and cost considerably. It is thus not surprising that they have gained such wide interest in recent years. Applications range from gene expression analysis over medical diagnosis to genetic fingerprinting and pathogen identification; for example, Fox (2000) describes an assay to identify bacteria contaminating drinking water; (Delpech, 2000) describes applications in diagnostics in medicine.

For such DNA array experiments to succeed, appropriate oligonucleotide probes have to be selected for each of the sequences to be identified, i.e. for each individual spot on the array surface. Given a set of genomic sequences, called *target sequences* in the following, the objective is to find one oligonucleotide (called *probe* here) for each target sequence in the set. These probes will then be attached to the array surface. Each probe on the array should hybridize only to the intended target, and not to any other sequence in the target set, i.e. a probe must have a high specificity in detecting the target. The problem is further complicated as all probes must work under the same hybridization conditions, most importantly, at the same temperature. The problem can be formalized as follows: Given n target sequences $t_1, t_2, \ldots, t_n$, find a temperature T and n probe sequences $p_1, p_2, \ldots, p_n$ such that

$$T_M(p_i, t_i) - \epsilon > T > T_M(p_i, t_k) + \epsilon \qquad (1)$$

for all $k \neq i, i = 1, \ldots, n$, where $T_M(x, y)$ is the temperature below which the two strands x and y are bound, and above which they denature. T is the temperature at which the experiment should be carried out. The additional temperature margin ϵ compensates for example for model errors and imprecisions.

Melting theory and nearest neighbor model

The computation of T_M for a given duplex is based on the assumption that we deal with two-state transitions: Either the DNA is in the double helical state, or it is in the random coil, denatured state. Clearly, this

[*] To whom correspondence should be addressed.

© Oxford University Press 2002

presents a simplification. Synthetic polymers with simple repeat sequences usually melt in a single cooperative transition, however, natural polymers with heterogeneous sequences may melt with many stable intermediate states, and accurate predictions require a statistical mechanical partition function approach (SantaLucia, 1998). Such an approach, however, is too complex for a full all-against-all melting temperature computation as we suggest for probe design, and we assume that a two-state model will reasonably well approximate the true melting behavior of short oligonucleotides as used on DNA chips.

We consider the two-state reversible equilibrium annealing reaction of two DNA single strands (compare (Owczarzy *et al.*, 1997))

$$S_1 + S_2 \overset{K_D}{\rightleftharpoons} D \tag{2}$$

where K_D is the equilibrium constant.

T_M is defined as the temperature at which 50% of the strands are in the double stranded and 50% in the random coil, denatured state. It can be shown that (cf. Freier *et al.*, 1986; Ornstein and Fresco, 1983; Owczarzy *et al.*, 1997; Rychlik and Rhoads, 1989):

$$T_M = \frac{\Delta H}{\Delta S + R \ln C_T/4}, \tag{3}$$

where ΔH and ΔS are enthalpy and entropy changes of the nucleation reaction, R is the Boltzmann constant, and $C_T = [S_1] + [S_2] + 2[D]$ is the total molar concentration of strands.

This concentration dependence of T_M induces some problems to our ansatz of calculation, as target DNA concentration is unknown in DNA array experiments. Thus the calculation cannot be accurate. However, (Li and Stormo, 2001) report that T_M is still sufficiently precise for probe evaluation. They suggest using a constant of $1 \times 10^{-6} M$ for C_T.

Interactions between bases in nucleic acids are of two kinds (Cupal, 1997):

- *Base pairing* in the plane of the bases due to hydrogen bonding between base pairs in the two opposing strands, and
- *Base stacking* perpendicular to the plane of the bases due to London dispersion forces and hydrophobic effects.

Both quantum chemical calculations and thermodynamic measurements suggest that base pairing contributions to total energy depend exclusively on base pair composition, while stacking contributions depend on base pair composition and base sequence along the chain. Obviously, models based solely on base composition neglect stacking contributions, and yield less precise results (Rychlik and Rhoads, 1989).

As the major contribution to the overall stabilizing energy of nucleic acid structures results from short-range interactions, we assume that the stability of a base pair (and its contribution to enthalpy and entropy of the duplex) depends only on the identity of its immediate up- and downstream neighbors. This assumption leads to the Nearest Neighbor (NN) Model. In this model one assumes that ΔH and ΔS of the melting reaction can be calculated by summing up the contributions of the individual neighboring pairs. ΔH and ΔS can then be used with Equation (3) to calculate the melting temperature of the strands.

Usually, thermodynamic parameters for the nearest neighbor model are determined from UV-absorbance vs temperature profiles of a number of different, short oligonucleotides. By fitting the measured curves to the model, parameters can be obtained that according to SantaLucia on average fit ΔG, ΔH, ΔS, and T_M within 4%, 7%, 8% and 2 degrees Celsius, respectively (SantaLucia *et al.*, 1996). Parameters are available for DNA–DNA (Allawi and SantaLucia, 1997, 1998a,b,c; Breslauer *et al.*, 1986; Gotoh and Tagashira, 1981; Peyret *et al.*, 1999; Quartin and Wetmur, 1989; SantaLucia *et al.*, 1996; SantaLucia, 1998; Sugimoto *et al.*, 1996), RNA–RNA (Freier *et al.*, 1986; SantaLucia and Turner, 1997; Xia *et al.*, 1998) and DNA-RNA (Gray, 1997) duplexes, a number of those with additional corrective factors to adjust for nonstandard hybridization conditions.

ALGORITHM

Probes should bind specifically to the target sequences. Therefore, Kurtz *et al.* (2001) suggest using probes that are complementary to their respective target sequence, and that are unique up to k errors. They use a suffix tree to identify such unique candidates, and suggest using other software to select the actual probes to be used in the experiment. Similarly, we consider only probes that are perfectly complementary to their respective target sequence, and that are unique. These probes are then evaluated further using the Nearest Neighbor Thermodynamic Model. It should be noted that in some cases admitting probes with mismatches can increase specificity if the target sequences are very similar. Our algorithm will not consider such probes for complexity reasons.

To select optimum complementary probes, melting temperatures between the complements of all substrings of all target sequences (the probe candidates) and all targets have to be computed. Application of the NN model requires knowledge about which basepairs are going to form in the duplex; hence an alignment of the probe and the target sequence is required, and the alignment resulting

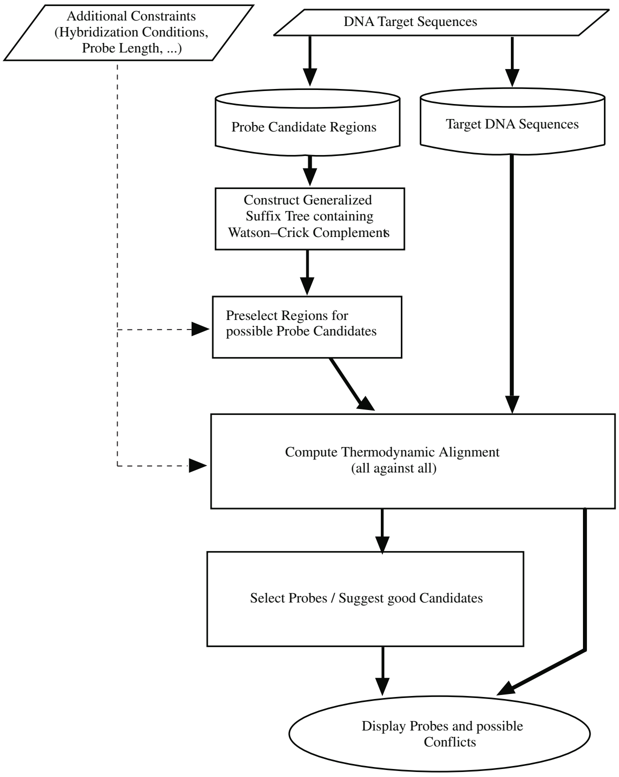

Fig. 1. Method overview for probe selection algorithm.

in the highest T_M is desired. Furthermore, as it is possible that the probe–target duplex involves more complex secondary structure, combinations containing mismatches and unpaired bases within the duplex (bulges, loops, etc.) will be considered as well. Note that foldback of a target sequence onto itself and target–target interactions can impede probe–target binding, and that consideration of such cases would require a native unimolecular folding algorithm such as MFOLD (Zuker, 1989) to be executed for all probe–target and target–target interactions. In the case of very complex target mixtures, such interactions may have an important effect on the outcome of hybridizations, but they can not be computed sufficiently efficient for our purposes.

Figure 1 gives an overview of the probe design algorithm. Given the DNA target sequences, our goal is to exclude infeasible probe candidates as early as possible. Infeasible probes are probes that are too short or too long, occur more than once in different targets, or do not fulfill other relevant criteria. Thermodynamic computations, which are quite time-consuming, will only be done for the remaining probe candidates.

The algorithm begins by constructing a generalized suffix tree ST from the inverse complements of all target sequences. A suffix tree is a data structure allowing for fast recognition of repetitive subsequences in strings. This property is used to identify non-unique probes, i.e. probes forming perfect duplexes with more than one target, which are subsequently removed from the probe candidate set. Also, some other criteria are used to remove infeasible probes, such as the probe length and the melting temperature of the probe with its Watson–Crick complement. We will come back to this point later. Note that additional criteria can be included easily.

Given the target DNA sequences and probe candidates, the algorithm computes melting temperatures for all combinations of probe–target interactions, i.e. melting temperatures between all probe candidates paired with all target sequences. As DNA is known to be highly repetitive (cf Gusfield, 1997, p.286) much time can be saved by avoiding recomputation of melting temperatures for subdomains of probes with some given target that have already been considered. The probes are stored in the generalized suffix tree ST in the preselection step, and this suffix tree is used further in the algorithm to avoid such redundant computations.

Finally, probes and melting temperatures are output, and oligos for the array can be chosen from suggestions made by our software.

Thermodynamic alignment

To apply the nearest neighbor model, we need to know which bases are going to form basepairs in the duplex. Unfortunately, this is not clear at all if the strands are not perfectly complementary to one another in the Watson–Crick sense. Worse yet, bases may remain unpaired within a duplex, and the duplex will still be quite stable (Ke and Wartell, 1995; LeBlanc and Morden, 1991; Turner, 1992). The problem is related to finding the minimum energy RNA secondary structure, which can be solved by a dynamic programming based algorithm similar to the one presented here (Zuker, 1989). The algorithm they present solves the folding problem exactly, also considering several more complex structures such as multiloops which our algorithm does not consider, however, at the expense of a considerable running time, making its adaptation infeasible for our application.

We compute an alignment of the two sequences, allowing for gaps. The alignment and T_M are interdependent: We cannot compute T_M without knowing the alignment, and the alignment should maximize T_M. Enumerating all possible alignments and computing their respective melting temperatures to choose the maximum thereof is infeasible, as the number of alignments grows exponentially with sequence length and the problem would quickly become computationally intractable.

The problem of aligning two sequences given a weight function $w(\cdot,\cdot)$ is one of the standard bioinformatics problems. A dynamic programming algorithm due to

Needleman–Wunsch (Durbin *et al.*, 1998; Gusfield, 1997; Waterman, 1995) can be used to find an alignment maximizing $w(\cdot, \cdot)$. The general idea is to consecutively extend the alignment, starting with an alignment of prefixes of the two sequences x and y.

We have modified this algorithm to calculate ΔH and ΔS for all prefix-alignments, choosing the one resulting in the highest local melting temperature: Our alignment cost function is the T_M function from Equation (3), storing ΔH and ΔS at every position in the table. Then, the dynamic programming recursion becomes

$$\Delta H_{i,j} = \begin{cases} \Delta H_{i-1,j-1} + \Delta\Delta H(x_i, y_j) & \text{if } t = 0 \\ \Delta H_{i-1,j} + \Delta\Delta H(x_i, -) & \text{if } t = 1 \\ \Delta H_{i,j-1} + \Delta\Delta H(-, y_j) & \text{if } t = 2 \end{cases} \tag{4}$$

$$\Delta S_{i,j} = \begin{cases} \Delta S_{i-1,j-1} + \Delta\Delta S(x_i, y_j) & \text{if } t = 0 \\ \Delta S_{i-1,j} + \Delta\Delta S(x_i, -) & \text{if } t = 1 \\ \Delta S_{i,j-1} + \Delta\Delta S(-, y_j) & \text{if } t = 2 \end{cases} \tag{5}$$

and $t \in \{0, 1, 2\}$ is to be chosen such that

$$T_M(i, j) = \frac{\Delta H_{i,j}}{\Delta S_{i,j} + R \ln C_T/4} \tag{6}$$

is maximal. Note, that $\Delta\Delta H(x_i, y_j)$ and $\Delta\Delta S(x_i, y_j)$ denote the nearest neighbor parameters for enthalpy and entropy changes, respectively, when the ith base of x, x_i and the jth base of y, y_j are paired in the alignment. '–' stands for a gap in the alignment, representing an unpaired base in the duplex. Note also, that $\Delta\Delta H$ and $\Delta\Delta S$ depend not only on the current basepair, but also on the one before (the nearest neighbor). However, implementing this dependency is straightforward. We neglect this issue here for the sake of simplicity.

By initializing the border of the dynamic programming table with zeros, we assure that initial gaps do not lower T_M; by looking for the result not just in cell $(|x|, |y|)$, but in cells $(s, |y|)$ and $(|x|, t)$ for all $s = 1..|x|$ and $t = 1..|y|$ and choosing the maximum value found, the same is true for terminal gaps. Furthermore, a special symbol is used to denote the beginning and the end of a sequence, allowing the use of dangling end thermodynamic parameters in the computation (Bommarito *et al.*, 2000).

Unfortunately, the melting temperature Equation (6) does not show strict monotonicity, which can cause the the alignment algorithm to return a suboptimal alignment in some cases. An example is given below. To assess the quality of the approximation using the alignment algorithm with parameters listed in Kaderali (2001), we have enumerated all perfect Watson-Crick duplexes of length up to 15 nucleotides, and shown that the algorithm finds the optimum alignment in all these cases. Furthermore, for over 100,000 random Watson–Crick duplexes of length up to 250, not a single error was made either. In the case where the most stable duplex contains one single unpaired nucleotide, the greedy approach may fail. Consider, for example, the duplex

```
0 1 2 3 4 5 6 7 8 9
G T G T G C A A A A
- C C A C G T T T T
. M M M M M M M M M
```

with a melting temperature $T_M = 27.2$°C, whereas the alignment algorithm finds

```
0 1 2 3 4 5 6 7 8 9
G T G T G C A A A A
C - C A C G T T T T
M . M M M M M M M M
```

with $T_M = 15.4$°C. The difference is caused when the algorithm is forming the G/C pair in position 2. It has to decide between either the GT/C- alignment or the GT/-C alignment. The alignment resulting in the higher local melting temperature is chosen—but unfortunately, when more bases are added after the G/C pair, it turns out that the wrong choice has been made.

To be able to estimate the magnitude of the error, two Monte Carlo computer experiments have been made:

(1) Generate two random sequences of random lengths between *minlen* and *maxlen* nucleotides; note that the two sequences generated may be of different length. The nucleotides in each sequence are drawn independently and from an identical, fixed distribution. Run the thermodynamic alignment algorithm to calculate the alignment melting temperature T_M^{align}. In parallel, enumerate all possible alignments, calculate their respective melting temperatures, and save the maximum T_M^{enum} thereof.

(2) Generate one random sequence of random length between *minlen* and *maxlen*, using the same procedure as above. Then construct a second sequence as the Watson–Crick complement, and introduce at most *maxmut* insertions, deletions or substitutions. Again, run the thermodynamic alignment algorithm to calculate the alignment melting temperature T_M^{align}. In parallel, enumerate all possible alignments, calculate their respective melting temperatures, and save the maximum T_M^{enum} thereof.

Computer experiment 1 has been carried out with $minlen = 10$ and $maxlen = 15$ for 2500 random sequences. The results are reassuring. Figure 2 depicts the difference $T_M^{enum} - T_M^{align}$ between the melting temperature of the optimum alignment and the solution found by our algorithm (rounded **up** to the next integer), the bars showing how many of the 2500 computations had an error

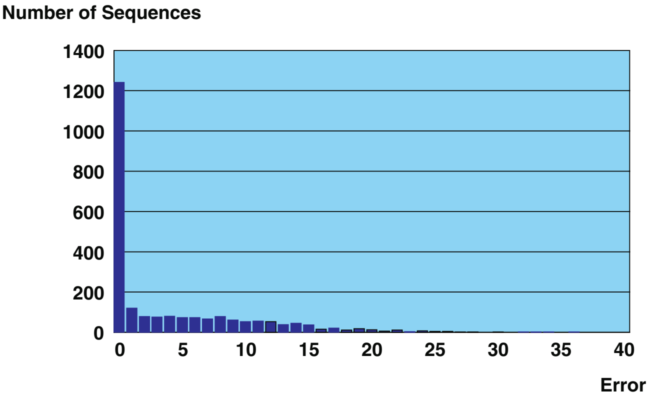

Fig. 2. The diagram shows the error made by the algorithm for alignments of random sequences of lengths between 10 to 15 nucleotides, the bar indicating how many of the 2500 alignments calculated showed an error of the respective magnitude.

of the respective magnitude. The average error made was 3.13 degrees Celsius, the maximum error observed was 35.46 degrees.

The case where the calculated temperature is too low when the sequences forming the duplex have only little similarity is not really a problem—the actual melting temperature of the duplex will be too low to play a role in probe design anyway. Therefore, the results of experiment 2 are even more interesting. Again, the experiment was conducted for 2500 alignments with at most one mutation. In that case, no error was made in 87.12% of the cases. An error of not more than three degrees was made in 90.12% of the alignments. The average error made was 1.27°C, the maximum error was 34.3 degrees Celsius.

Suffix trees

The algorithmic idea introduced in this section will reduce running time further. Nothing is lost in terms of result quality, all we need is a little more memory and an additional data structure.

The underlying idea is straightforward. Assume we have just computed the dynamic programming table for the two sequences 'GATTACA' and 'CTAAGGT'. Further assume we need to align 'GATTACA' and 'CTAATGA' sometime thereafter. Then, the two dynamic programming tables share the subtable for 'GATTACA' and 'CTAA'. We need not recompute this part of the dynamic programming table for the latter alignment, but may use the subtable from the former alignment and compute only the remaining, different entries. To identify common prefixes of substring pairs of *all* the different sequences under consideration, we use a generalized suffix tree. Note, that the same tree is used in both probe preselection and alignment.

A suffix tree for the sequence 'TACTACA' is shown in Figure 3. Its defining property is that each path from the root node to a leaf corresponds to a suffix of the string represented by the tree, and vice versa. Note that, by appending the unique character '$' at the end of the string, we guarantee that every suffix ends at a leaf. Otherwise, the suffix 'A' of 'TACTACA', i.e. the suffix consisting of only the last character of the original sequence, would end within the 'AC' edge. This problem arises whenever a suffix of the string is a prefix of another suffix.

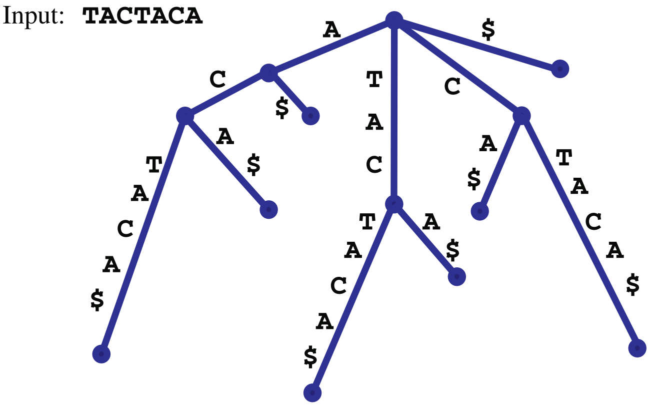

Fig. 3. Suffix tree for the sequence 'TACTACA'. Note how all the suffixes 'TACTACA', 'ACTACA', 'CTACA', 'TACA', 'ACA', 'CA', 'A' and the empty suffix are described by a unique path from the root node to one of the leaves, and how every leaf uniquely yields one such suffix. The symbol '$' denotes the end of a string.

Suffix trees can be constructed in linear time in the length of the given string. Esko Ukkonen (Ukkonen, 1995) devised a straightforward $\mathcal{O}(n)$ algorithm in 1995. An excellent description of that algorithm can be found in Gusfield (1997).

A *Generalized Suffix Tree* is a suffix tree containing all suffixes of a finite number of strings. Only slight modifications are required to construct generalized suffix trees with Ukkonen's algorithm, and the resulting algorithm still runs in linear time.

Probe preselection

As mentioned above, we need to compute melting temperatures (and alignments) between all probe candidates and all sequences in the target set. Therefore, it seems worthwhile to put some effort into reducing the number of probe candidates before doing so. There are several criteria that help exclude infeasible probes:

- *Probe Length:* Usually, there are some restrictions to probe length. These may be due to technical limitations in the process of array manufacturing, as well as limitations given by the user or other external causes. Without going into more detail at this point, we assume that we have variables $minlen$ and $maxlen$ with $minlen \leq |probe_i| \leq maxlen$, where $|probe_i|$ is the length of probe i, and all feasible probes have to satisfy that inequality.

- *Unique Probes:* If a given probe is the perfect Watson–Crick complement to substrings of two or more target sequences, that probe will hybridize to both targets with the same melting temperature. Therefore, such probes cannot be used for array experiments, as both targets would hybridize against the same spot on the array. We will allow only probes that are complementary to exactly one substring of all target sequences. Furthermore, a probe can even be excluded if it contains a substring longer than some given length l (say, 80%), that is complementary to more than one substring of all target sequences. Such substrings can easily be identified using hashing-techniques.
- *Probe Melting Temperature:* Last but not least, one can impose some constraints on the minimum temperature that a probe–target duplex should be able to withstand. The array experiment will be carried out at some temperature T, therefore $T_M(target, probe) > T$ must hold. Of course, the problem of determining T and the probes to be used are not independent from one another. However, we assume some bound $T_B \leq T$ to be given that can be used to exclude probes with $T_M(probe, target) < T_B$ from further consideration.

The algorithm to preselect probes starts with the set of complements of all substrings of all the target sequences. Every substring of a string is a prefix of a suffix of that string. Therefore, a generalized suffix tree can be used to represent all substrings. By following a unique path from the root to another node, either leaf or internal, all substrings can be retrieved from the tree (the path may end somewhere within an edge, i.e. it need not necessarily terminate at a node). Similarly, each substring corresponds to one such path starting at the root node. Note, however, that one path may correspond to two (equal) substrings from the same or different sequences.

After the generalized suffix tree ST containing the complements of all target sequences has been constructed, applying the above criteria and removing all infeasible probes from the tree is straightforward. This pruning of the tree yields the suffix tree ST_{pruned}, which is then used further in the following steps.

Thermodynamic tree alignment

Recall, that our objective is to determine an oligo probe for each of the n genomic sequences that will hybridize only to its respective target, and not to any of the other sequences. Hence, we need to compute the melting temperatures of the most stable duplex formed between each probe left after preprocessing and each target. The final step of selecting probes from the output of the thermodynamic tree alignment algorithm will be described in the next section.

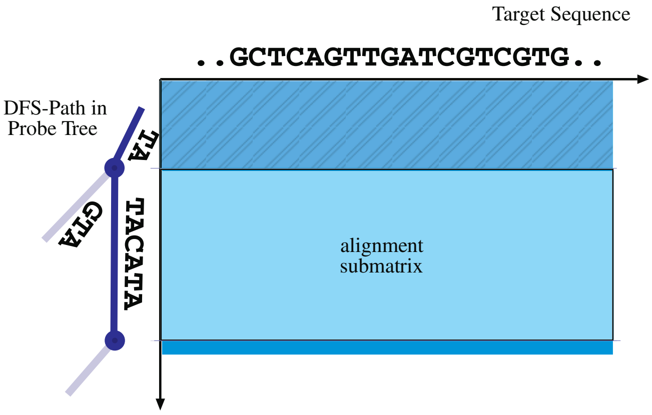

Fig. 4. Alignment of the target sequence with probe ..$TATACATA$... Note that if the alignment of the sequence TAGTA and the same target has been computed before, the upper part of the dynamic programming table can be reused.

To compute the melting temperatures, begin with the complements of all substrings of the n target sequences. This is done by constructing generalized suffix tree ST containing the complements of all target sequences, as described above. Then, all substrings are contained in that tree. The second step is to reduce the number of substrings stored in the tree, this is taken care of as described in the previous section on probe preselection, yielding the pruned tree ST_{pruned}.

Finally, all that remains to be done is the computation of the melting temperatures of the duplexes formed between all substrings left in the tree and all target sequences; i.e. each substring and each target have to be aligned using the thermodynamic alignment algorithm described above, and the maximum melting temperature must be determined. Doing so is extremely time-consuming. We will therefore use the pruned suffix tree ST_{pruned} from the preprocessing step to reduce running time.

Repetitive subsequences in DNA are quite common. Thus, whenever calculating alignments of two strings, we may be able to reuse parts of the dynamic programming table from a previously computed alignment, if the strings from that prior alignment share prefixes with the actual strings.

Fortunately, we can use the tree ST_{pruned} constructed during preprocessing to identify such common prefixes of probes. The tree induces an ordering of the probes, grouping probes with common prefixes together. This helps to calculate such groups at a time and to avoid the storage of different subtables, which reduces the overall memory requirements of the program. Our empirical results as well as some theoretical considerations indicate that the suffix tree increases the computation speed by a factor of around five. Note that each path in the suffix tree is aligned against all target sequences, but the target sequences are not stored in a tree. Obviously, storing the target sequences in a second tree and aligning the two trees could result in additional speed gains. We are currently investigating this idea further.

A second trick introduced to speed up the alignment computation is to check, if the two sequences under consideration share a complementary substring-pair of at least some given length k. If this is not the case, we assume that the resulting melting temperature will be very low, and skip its computation.

Oligo probe selection

The thermodynamic tree alignment algorithm determines probe candidates according to certain criteria, and returns melting temperatures $T_M(probe, target)$ for all $(probe\ candidate, target)$ pairs, i.e. all probe candidates and all target sequences.

Given the output list from the thermodynamic tree alignment algorithm, our objective now is to select a temperature T and one probe from the list for each of the target sequences, such that the array experiment can be carried out at temperature T, and the probes selected will hybridize only to their intended target sequence, and not to any of the other sequences. This problem can be formalized as follows:

Given n DNA or RNA target sequences $t_1, t_2, .., t_n$, given furthermore for each target sequence t_i a finite set of probe sequences $\mathcal{P}_i$, where $\mathcal{P}_i \bigcap \mathcal{P}_j = \emptyset$ for all $i, j; i \neq j$. Furthermore given for all target sequences t_i and all probe candidates $p_j \in \bigcup_{k=1}^{n} \mathcal{P}_k$ the melting temperatures $T_M(t_i, p_j)$ at which target t_i and probe p_j dissociate.

Find a temperature T and, for each target sequence t_i, select one probe $p_k \in \mathcal{P}_i$ s.t.

$$T_M(t_i, p_k) \geq T > T_M(t_j, p_k) \qquad (7)$$

for all $j \neq i$.

The temperature T is a temperature that must hold for all probes selected; the inequality above must be satisfied by all probes selected for all targets with the same temperature T. This implies that for two selected probes p_i for target t_i and p_j for target t_j, the inequalities $T_M(t_i, p_i) > T_M(t_j, p_i)$, $T_M(t_i, p_i) > T_M(t_i, p_j)$, $T_M(t_j, p_j) > T_M(t_i, p_j)$ and $T_M(t_j, p_j) > T_M(t_j, p_i)$ must hold: All ‘desired’ hybridizations have melting temperatures higher than all ‘undesired’ cross-hybridizations.

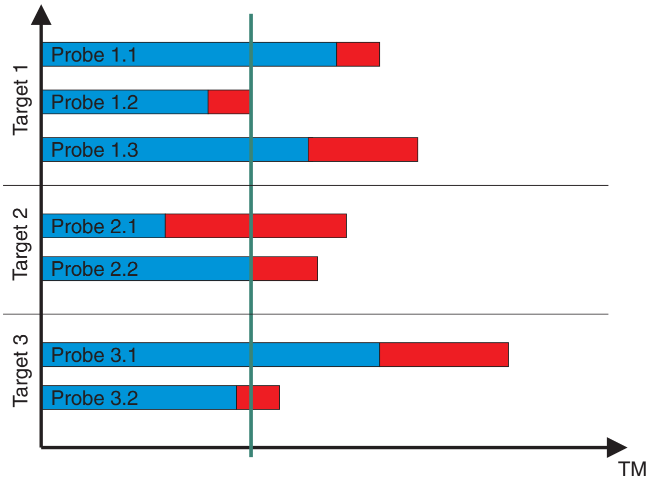

Fig. 5. Probe selection. The X-axis represents the melting temperature, the Y-axis the different probes. Probe indices are of the form $t.p$, where t refers to the target sequence the respective probe is intended for, and p is the number of the probe for that target. For each probe, the right end of the dark bar shows T_M of the perfect duplex formed between the probe and its respective target, the light gray bar shows the temperature range where the probe will crosshybridize. For the temperature represented by the line vertically crossing all probes, probes 1.2, 2.1 and 3.2 would yield a feasible set of probes. The algorithm starts with the maximum melting temperature found for some duplex, and decreases T until such a set is found.

This problem can be solved in polynomial time. The idea is to sort the probes for each target according to their melting temperature. Then, starting with the highest temperature T, consecutively lower T, and remove all probes that will crosshybridize at the new temperature. This is iterated until either a feasible, unambiguous probe is found for every target, or until all probes have been removed. Figure 5 illustrates the procedure.

IMPLEMENTATION

The algorithms presented here have been implemented in C++. The *PROBESEL* program combines probe preselection and the thermodynamic alignment algorithm and calculates melting temperatures between probe candidates and all target sequences. The *Pickprb* program implements the Probe Selection Problem (PSP) Algorithm to select one probe for each target sequence from the output generated by *Probesel*.

The program code has been tested on Intel-PCs under Windows NT 4.0 with Microsoft’s Visual C++ 6.0, on Sun Ultra Enterprise 4000 running Solaris 7 with the GNU g++ compiler, and on DEC Alpha / Compaq Tru64 UNIX V5.1 with Compaq’s cxx compiler, version 6.20.

DISCUSSION

Identifying HIV-1 subtypes

Besides running on randomly generated sequences of different lengths, the algorithm has been used to find oligos to be used for the identification of different HIV-1 subtypes. The complete HIV-1 reference subtypes database from Los Alamos National Laboratory, USA (LANL, 1999) has been processed. This database contains 58 sequences of average length around 9300 nucleotides. All oligos of length between 19 and 21 nucleotides with a melting temperature above 70°C have been evaluated. The parameter l (maximum permissible length of consecutive Watson–Crick basepairs in an unintended duplex) was set to 12, i.e. probes sharing a reverse-complemented subsequence of length > 12 bases with false binding sites were excluded from the probe set, and T_M was considered negligible for duplexes with less than ten consecutive basepairs (parameter k), i.e. for those, no alignment was calculated. The entire computation took 61 minutes for the *Probesel* program, and only a couple of seconds for *Pickprb* on a Compaq Tru64 machine with four DEC Alpha EV6.7 processors each operating at 667 MHz, and equipped with an alpha internal floating point processor. Note that the present version of the program runs single-threaded and hence makes no use of the multiple processors available.

Probes were found for all 58 sequences, with melting temperatures between 73° and 87° C. The highest temperature for which crosshybridizations are predicted is 53° C, which gives a margin of 20°. The program suggests conducting the experiment at a temperature of 63° C. Both the input file and the probes selected by the algorithm are available from our website for review.

Application to 28S rDNA sequences

The algorithms presented here have been applied to a database of 1230 28S rDNA sequences from different organisms (Markmann, 2000). Those 1230 sequences are of length between 160 and 6198 bases, with an average length of 676 nucleotides. As the database contains sequences with very high similarity ($> 95\%$), it was filtered before starting the *Probesel* program. To do so, pairwise Smith–Waterman alignments of all sequences were computed, using edit distance as distance function. Then, for each aligned pair of sequences, all matches between the two sequences were counted. This was set in relation to the length of each of the sequences in the alignment, including internal gaps, but not counting initial and terminal gaps. Whenever some sequence was over 95% similar to another sequence according to that measure, it was removed. If both sequences had relative similarity of over 95% to one another, the shorter one was removed.

487 sequences remained in the database after this preprocessing step. Then, the *Probesel* algorithm was started with probe length 29–30 and minimum probe-target melting temperature 60° C. No unique probes could be found for 44 sequences, which *Probesel* reported after approximately 2 minutes.

For $k = 10$ and $l = 12$, the computation took less than 1 hour, however, no probes could be found for 186 of the 443 target sequences when requiring a distance of 10 degrees between the lowest temperature for intended hybridizations and the highest temperature for crosshybridizations. For the remaining 257 sequences, the algorithm suggests to conduct the experiment at a temperature of 60 degrees.

To improve on that result, the program was run again, this time with $k = 0$ and $t = 15$, i.e. no probes were excluded based on consecutive basepairs formed with unintended targets. Under those circumstances, the computation took considerably longer, approximately 7 hours. Therefore, probes were found for all but 46 targets, with a temperature margin of 10 degrees. The algorithm suggests washing the array at 66 degrees. If a temperature margin of only 5 degrees is required, oligo probes were found for all but 35 of the 443 sequences.

Experimental evaluation

To evaluate the sensitivity and specificity of the crosshybridization predictions made by our software, predicted pairwise melting temperatures were compared against hybridizations observed in an experiment conducted at Los Alamos National Laboratory. In this experiment, 25 oligonucleotides (tags) and their complements (anti-tags) were synthesized. The tags were attached to plastic beads, and the anti-tags flourescently labeled. Then, one anti-tag was brought to reaction with all tags at a time, and the results were read using a flow cytometer. Figure 6 shows the results. All pairwise interactions were then simulated *in silico* using our melting temperature algorithm, the results are also shown in the figure (all reactions with TM above 60 degrees shown). Note that the crossreactions for tag S7 with all other anti-tags have been confirmed to be experimental artefacts.

Experimental evaluation of the probes selected by the algorithm for the 28S rDNA application is under way. Results will be published elsewhere.

Outlook

First results of using the algorithm to find probes for all yeast ORFs indicate that it is feasible to apply this approach to genome-scale applications. Both plus and minus strand of the whole yeast genome were evaluated, resulting in 6165 target sequences after filtering as described in the 28S rDNA case. The program ran for about 2 weeks. This is clearly acceptable, especially

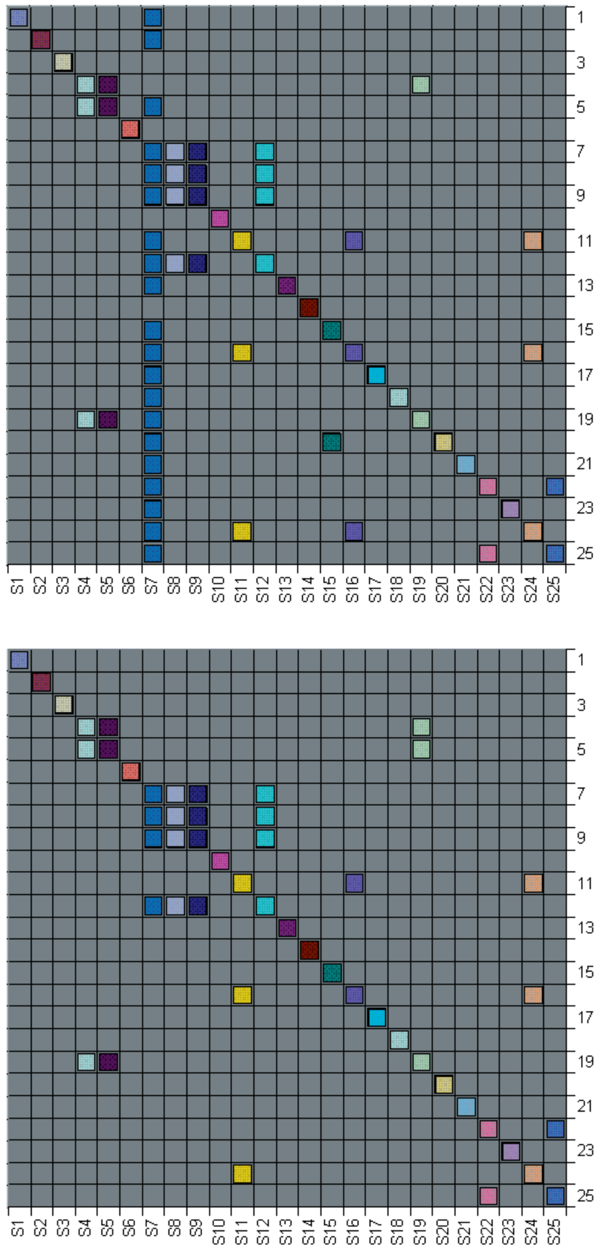

Fig. 6. This plot shows experimentally determined (top) and calculated (bottom) crossreactions for 25 oligonucleotides of length 20.

when considering that the algorithm could easily be parallelized, and the software is currently not optimized for speed. Probes were found for 4431 of the ORFs, with an experimental temperature of 65 degrees. Note, that careful selection of filtering criteria and choice of parameters can greatly affect running time, which makes a complexity analysis of the algorithm extremely difficult and which we therefore cannot provide at this time.

Already the 28S rDNA example reported above indicates that it is very hard to find oligonucleotides for larger datasets. This problem becomes much worse for genome-scale applications, for which we suggest three solutions: Either use longer probes, adding to the difficulties and costs of manufacturing the array; conduct several experiments at different temperatures, each experiment testing a subset of the target sequences; or attach several different short oligonucleotides for the same target to the array surface. Such oligonucleotides will show crosshybridization. However, provided they are appropriately chosen, it is possible to derive the expression level for each target strand. (Knill *et al.*, 1996) solve a similar problem using Markov chain Monte Carlo methods; further research following this idea is currently being conducted at the University of Cologne.

ACKNOWLEDGEMENTS

We would like to thank Diethard Tautz, Institute of Genetics, University of Cologne for pointing out the problem and helping with various questions concerning biology, biochemistry and genetics. Alexander Pozhitkov contributed greatly by his readiness to spend hours discussing hybridization properties of nucleic acids, and Melanie Markmann provided her 28S rDNA database to test the algorithm.

We are especially grateful to P. Scott White and Francisco J.Uribe, Bioscience Division (B-1), Los Alamos National Laboratory, for providing their experimental data to test the predictions made by the algorithm, and for their helpful comments and suggestions on both the paper and the software, and we gratefully acknowledge the helpful remarks of the anonymous referees.

REFERENCES

Allawi,H. and SantaLucia,J. (1997) Thermodynamics and NMR of internal g-t mismatches in DNA. *Biochemistry*, **36**, 10581–10594.

Allawi,H. and SantaLucia,J. (1998a) Nearest neighbor parameters for internal g-a mismatches in DNA. *Biochemistry*, **37**, 2170–2179.

Allawi,H. and SantaLucia,J. (1998b) Nearest neighbor parameters of internal a-c mismatches in DNA: sequence dependence and pH effects. *Biochemistry*, **37**, 9435–9444.

Allawi,H. and SantaLucia,J. (1998c) Thermodynamics and NMR of internal c-t mismatches in DNA. *Nucleic Acids Res.*, **26**, 2694–2701.

Bommarito,S., Peyret,N. and SantaLucia Jr,J. (2000) Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res.*, **28**, 1929–1934.

Breslauer,K., Frank,R., Blöcker,H. and Marky,L. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.

Cupal,J. (1997) The density of states of RNA secondary structures, Master's thesis, University of Vienna.

Delpech,M. (2000) Les puces à ADN. *Annales de Biologie Clinique*, **58**, 29–38.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Fox,G.E. (2000) Rapid identification of unexpected bacteria pathogens in space environments. http://www.isso.uh.edu/publications/A9900/html/mini/mini-fox.htm.

Freier,S., Kierzek,R., Jaeger,J., Sugimoto,N., Caruthers,M., Neilson,T. and Turner,D. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl Acad. Sci. USA*, **83**, 9373–9377.

Gotoh,O. and Tagashira,Y. (1981) Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles. *Biopolymers*, **20**, 1033–1042.

Gray,D. (1997) Derivation of nearest-neighbor properties from data on nucleic acid oligomers. II. thermodynamic parameters of DNA–RNA hybrids and DNA duplexes. *Biopolymers*, **42**, 795–810.

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

Kaderali,L. (2001) Selecting target specific probes for DNA arrays, Master's thesis, University of Cologne.

Ke,S. and Wartell,R. (1995) Influence of neighboring base pairs on the stability of single base bulges and base pairs in a DNA fragment. *Biochemistry*, **34**, 4593–4600.

Knill,E.H., Schliep,A. and Torney,D.C. (1996) Interpretation of pooling experiments using the Markov chain Monte Carlo method. *J. Comput. Biol.*, **3**, 395–406.

Kurtz,S., Choudhuri,J., Ohlebusch,E., Schleiermacher,C., Stoye,J. and Giegerich,R. (2001) Reputer: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, **29**, 4633–4642.

LANL, (1999) HIV sequence database - 1999 HIV-1 subtype reference alignments. Los Alamos National Laboratories,http://hiv-web.lanl.gov/.

LeBlanc,D. and Morden,K. (1991) Thermodynamic characterization of deoxyribooligonucleotide duplexes containing bulges. *Biochemistry*, **30**, 4042–4047.

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

Markmann,M. (2000) Entwicklung und Anwendung einer 28S rDNA-Sequenzdatenbank zur Aufschlüsselung der Artenvielfalt limnischer Meiobenthosfauna im Hinblick auf den Einsatz moderner Chiptechnologie, PhD thesis, University of Munich.

Ornstein,R. and Fresco,J. (1983) Correlation of t_m and sequence of DNA duplexes with δh computed by an improved empirical potential method. *Biopolymers*, **22**, 1979–2000.

Owczarzy,R., Vallone,P., Gallo,F., Paner,T., Lane,M. and Benight,A. (1997) Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers*, **44**, 217–239.

Peyret,N., Seneviratne,P., Allawi,H. and SantaLucia,J. (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.

Quartin,R. and Wetmur,J. (1989) Effect of ionic strength on the hybridization of oligonucleotides with reduced charge due to methylphosponate linkages to unmodified oligodeoxynucleotides containing the complementary sequence. *Biochemistry*, **28**, 1040–1047.

Rychlik,W. and Rhoads,R. (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and *in vitro* amplification of DNA. *Nucleic Acids Res.*, **17**, 8543–8551.

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

SantaLucia Jr,J., Allawi,H. and Seneviratne,P. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.

SantaLucia Jr.,J. and Turner,D. (1997) Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers*, **44**, 309–319.

Sugimoto,N., Nakano,S., Yoneyama,M. and Honda,K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.*, **24**, 4501–4505.

Turner,D. (1992) Bulges in nucleic acids. *Curr. Opin. Struct. Biol.*, **2**, 334–337.

Ukkonen,E. (1995) On-line construction of suffix-trees. *Algorithmica*, **14**, 249–260.

Waterman,M. (1995) *Introduction to Computational Biology*. Cambridge University Press, Cambridge.

Xia,T., SantaLucia,J., Burkard,M., Kierzyk,R., Schroeder,S., Jiao,X., Cox,C. and Turner,D. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry*, **37**, 14719–14735.

Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.