BIOINFORMATICS

Vol. 16 no. 10 2000
Pages 890–898

Information theoretical probe selection for hybridisation experiments

Ralf Herwig [1,*], Armin O. Schmitt [2], Matthias Steinfath [1], John O'Brien [1], Henrik Seidel [3], Sebastian Meier-Ewert [4], Hans Lehrach [1] and Uwe Radelof [1]

[1]Max-Planck Institut für Molekulare Genetik, Ihnestraße 73, D-14195 Berlin, [2]metaGen Gesellschaft für Genomforschung, Ihnestraße 63, D-14195 Berlin, [3]Schering AG, Genomics/Bioinformatics, D-13342 Berlin and [4]GPC AG, Fraunhoferstraße 20, D-82152 Martinsried, Germany

Received on March 21, 2000; accepted on May 23, 2000

Abstract

Motivation: *The choice of probes is an important feature of hybridisation experiments. In this paper we present an algorithm that optimises probes with respect to a training set of sequences based on Shannon entropy as a quality criterion. The practical motivation for our algorithm is oligonucleotide fingerprinting, a method for the simultaneous identification of sequences (cDNA or genomic DNA) by their hybridisation tags according to a set of short probes such as octamers, although the algorithm is of course not restricted to that application.*

Results: *We can show that our method is superior to the selection of probes according to their frequencies, which is a widely used strategy, and to randomly chosen probe sets. The quality of probe sets is assessed by a simulation pipeline that entails the set of probes as a simulation parameter. The performance of probe sets trained on sequences from different organisms shows additionally that probes should be chosen with regard to the organism under analysis. Case studies are presented on how constraints (*G + C*-content, complexity of the individual probes) influence the selection process.*

Availability: *A description of the oligonucleotide fingerprinting pipeline is published on our web-page http://www.molgen.mpg.de/~ag_onf/met.htm. An executable of the algorithm and probe lists designed for human and rodents can be downloaded from the ftp-site ftp://ftp.molgen.mpg.de/pub/mpimg/probe_design/.*

Contact: *herwig@molgen.mpg.de*

Introduction

Hybridisation experiments enable researchers to screen large amounts of clone material in parallel within one experiment, in contrast for example to gel fingerprinting techniques that involve separate handling of different clones, and are therefore well suited for large-scale genome analyses. The method of hybridisation of short synthetic oligonucleotide probes to cloned DNA sequences (cDNA clones or genomic DNA clones) in order to derive genetic sequence information has become a powerful tool in gene expression analysis (Poustka *et al.*, 1989; Lehrach *et al.*, 1990; Lennon and Lehrach, 1991; Meier-Ewert *et al.*, 1993; Maier *et al.*, 1994). Recent applications were published in the area of gene identification (Drmanac *et al.*, 1996; Milosavljevic *et al.*, 1996; Panopoulou *et al.*, 1998; Poustka *et al.*, 1999), where large numbers of cDNA clones in parallel were assigned to known genes or turned out to be candidates for new genes, in the field of comparative gene expression profiling (Meier-Ewert *et al.*, 1998), where cDNA libraries of different stages of development were compared and relevant development-specific genes were identified, and in the analysis of genomic DNA (Radelof *et al.*, 1998), where the method was used in order to reduce redundancy within shot-gun clone libraries. Besides scientific interest this technique has been established as an automated screening method in biotechnological industry in recent years.

PCR-products of a large number of cDNA or genomic clones are immobilised on nylon filter-membranes and hybridised in turn with 200–300 different, short (8-mers), radioactively labeled oligonucleotide probes of known sequences. Hybridisation signals are evaluated so that to each clone a vector of numerical values is assigned—its *fingerprint*—that represents the interaction of the clone sequence with the set of probe sequences. Detailed protocols, quality checks etc. of the procedure can be found in Schmitt *et al.* (1999) and Clark *et al.* (1999).

The crucial assumption of this approach is that the fingerprint is characteristic for the individual clone. For that reason the choice of probes is very important. The

*To whom correspondence should be addressed.

probes should be informative for the clone sequences in the sense that all different genes can be distinguished by their fingerprints. This implies that probes should occur within the clone sequences with a considerable frequency and that probes should not be too similar to each other. These requirements might point in different directions because the simple selection of probes according to high frequencies might lead to the agglomeration of probes that are highly similar to each other so that the gain in information about the clone sequences is not significantly increased when selecting probes successively.

Rather little attention has been paid on the proper design of probes in the literature. A method to design oligomers for hybridisation experiments was suggested by Cuticchia *et al.* (1993) in the context of physical mapping (see also Fu *et al.*, 1992). The criteria for the selection applied in that work were G + C-content of the oligomers combined with the expected frequency in a training database. The choice of probes according to frequencies is used also in Drmanac *et al.* (1996).

In this paper we discuss an information-theoretical approach to the design of probe sets that is based on entropy maximisation. Probes are calculated on the basis of a set of training sequences coming from (a) human and (b) rodent cDNA sequences. The performance of the probe sets is shown by evaluating a simulation pipeline that includes the probe sets as a simulation parameter. In both (a) and (b) we compare our method with the respective set of most frequent probes and a random collection and show that in both cases our choice is superior to any of the two alternatives. We also show that the performance of the probes is dependent on the training set and thus on the organism under analysis by comparing the performance of rodent- with human-trained probes on human test data. Additionally, we present some internal features of the algorithm when algorithmic parameters such as G + C-content and complexity of the probes are varied.

Systems and methods

Data preparation

In our practical applications we are mainly interested in the performance of the probes with respect to clustering sequences by evaluating pairwise similarities of their fingerprints. Clustering results depend—among other parameters—on the probe sets in use. Therefore, we implemented a data-analysis pipeline that contains the probe set as a simulation parameter. The final accuracy of the clustering result varies if fingerprints are derived from different probe sets so that clustering results can be used to assess the quality of the probe set and thus the quality of the algorithm. Figure 1 describes a flow chart of the procedure.

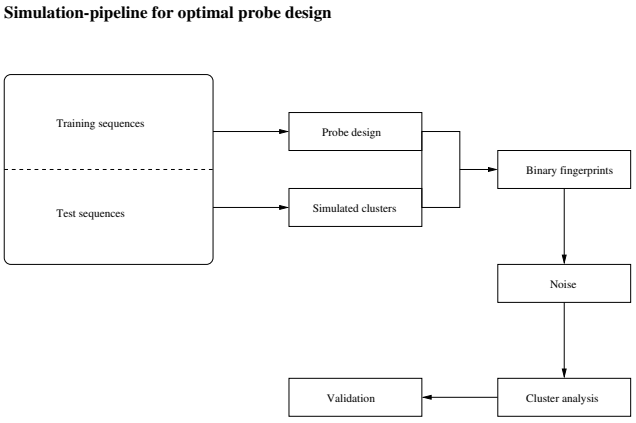

Fig. 1. Simulation pipeline for probe design. The set of probes enters the pipeline as a simulation parameter and influences the quality of the calculated clustering. By fixing all other parameters probe sets can be compared and judged according to the resulting clustering quality.

An initial set of sequences is divided into a part that is used as a training set and a part that is used to derive test sets for clustering. Probe sets are calculated from the training set using the different strategies and binary fingerprints are computed from the sequences of the test sets, i.e. sequences are replaced by binary fingerprints

$$\mathrm{ATGGTCATCCCGTC}\ldots = s_i \rightarrow f_i = \left(f_{i1}, \ldots, f_{ip}\right)$$

where f_i is the fingerprint of sequence s_i with respect to the p probes under analysis. $f_{ij} = 1$ if probe j or its reverse complementary sequence matches clone sequence i and 0 otherwise.

For each type of organism (human and rodent) we create five different test sets of 685 different cDNA sequences from the *GenBank* database. The sequences are copied according to a previously defined cluster distribution. Table 1 gives an overview on the simulated cluster structure. 500 sequences are present only one time (*singletons*), 144 sequences are copied 2–5 times, 18 sequences are copied 10–100 times and four sequences are present with more than 100 copies within the data set. The biggest of our test cluster has 201 members. Each test set results in a total of 2099 sequences.

Frequencies of cluster sizes are calculated by the formula $f(n) = \frac{s}{2n^2-n}$, where n is the size of the cluster, $f(n)$ is the number of different clusters of that size and s the number of singletons. The sizes of the bigger clusters are derived by a random number. This function reflects our experimental observations where we find most genes to be expressed with a low to moderate copy-rate (<50). We introduce noise within each data set via false positive rate, r_p, and false negative rate, r_n. Error

Table 1. Cluster distribution in simulated data sets. For example 144 cDNA sequences are copied 2–5 times which leads to a total of 388 copies that represent 18.5% of the data set

Copy rate	1	2–5	6–9	10–100	>100	Total
# sequences	500	144	19	18	4	685
# copies	500	388	136	401	674	2099
% of data set	23.8	18.5	6.5	19.1	32.1	100

is introduced independently for each probe by flipping the respective amount of digits of the binary fingerprints. Error parameters are constant in all data sets ($r_p = 0.2$, $r_n = 0.2$), i.e. 20% of the true positive signals are set to 0 and 20%, of the true negative signals are set to 1 for each probe in the fingerprint matrix (f_{ij}).

Clustering

The goal of the cluster analysis is to group the data around centroids so that similar data points are clustered together and dissimilar data points are separated; see Mirkin (1996) for a survey.

Our cluster analysis is performed with a sequential k-means algorithm using mutual information as a pairwise similarity measure for the binary fingerprints. This algorithm sequentially assigns each data point to the most similar cluster centroid from a set of previously calculated cluster centroids. The centroid is then updated by the data point and the next data point is processed. The algorithm is enriched by heuristics and algorithmic parameters that allow the merging of clusters and an introduction of new clusters in each step of the clustering process and thus does not need a pre-fixed initialization of the number of different clusters. A detailed description of the procedure and the similarity measure with applications within our simulation pipeline and with real data is described in Herwig *et al.* (1999). The algorithmic parameters remain unchanged for all simulation runs to focus results on the differing probe sets.

Validation of clustering

Assume a data structure of N data points, $x_1, \ldots, x_N$, where the true clustering, T, is known. Let $t_{ij} = 1$ if x_i and x_j belong to the same cluster and $t_{ij} = 0$ otherwise $(1 \leqslant i, j \leqslant N)$. For a calculated clustering, C, define similarly $c_{ij} = 1$ if x_i and x_j belong to the same cluster and $c_{ij} = 0$ otherwise, $(1 \leqslant i, j \leqslant N)$.

Validation of clustering is done by assigning each calculated clustering a numerical value that represents its quality with respect to the true clustering. To measure clustering quality we evaluate the 2×2 contingency table of the following form:

	0	1	Total
0	N_{00}	N_{01}	$N_{0.}$
1	N_{10}	N_{11}	$N_{1.}$
Total	$N_{.0}$	$N_{.1}$	$N_{..}$

where $N_{kl} = \#\{(i, j) ; t_{ij} = l, c_{ij} = k, 1 \leqslant i, j \leqslant N\}$, $0 \leqslant k, l \leqslant 1$, and where $N_{.k}$ and $N_{l.}$ are the respective marginal frequencies. Here the rows of the contingency table correspond to the calculated clustering and columns correspond to the true clustering. Clearly $N_{..} = N^2$. As a method of comparison we use the Jaccard-coefficient:

$$J(C, T) = \frac{N_{11}}{N_{01} + N_{10} + N_{11}}.$$

This measure takes into account only pairs that are clustered together, it does not value pairs that are not clustered together in either of the clusterings. This is advantageous because the high number of pairs that are not clustered together might dominate the quality measure. The range of the Jaccard-coefficient lies within the interval [0,1], where $J(C, T) = 1$ denotes a perfect clustering.

This measurement allows judging the quality of clustering and—when fixing all other parameters of the pipeline—judging the quality of the probe set under analysis.

Algorithm and implementation

The algorithm presented here has been implemented in a computer program written in the C programming language. It has been compiled with the Gnu-compiler and has been tested and run on Digital-Alpha-workstations under the *Digital Unix* operating system.

Assume a set of N sequences, $s_1, \ldots, s_N$. The fingerprints obtained with a single probe generate a partitioning of the N sequences into two subsets, i.e. those sequences that match with the probe sequence or its reverse complementary sequence and those that do not. The amount of information of the probe with respect to the set of sequences can be measured by entropy as introduced by Shannon (1948) (see also Cover and Thomas, 1991, for a survey):

$$I = -\sum_i p_i \log_2 p_i$$

where p_i is the proportion of sequences that fall in the respective subset, $i = 1, 2$. The entropy (or information content) is maximised when the subsets are of equal sizes, i.e. $p_1 = p_2 = 0.5$, so that a probe with a matching rate closest to 50% would lead to the highest information content about the sequences. Two probes should be used

so that the total entropy of the resulting four subsets of possible fingerprints is maximised, etc.

Because the number of possible fingerprints increases as 2^p with the number p of probes, screening all possibilities is computationally unfeasible. Therefore we use the following approximation which has originally been suggested by R.Mott (in Meier-Ewert, 1994):

(1) Find the probe which partitions best the set of known sequences (training set) into two groups (hybridising or not with that probe) which should be as equal in size as possible.

(2) Find the second probe which, together with the previously selected one, partitions the training set into four groups, which should be as equal in size as possible.

(3) In general, find the probe, which together with the previously selected ones, partitions best the training set.

(4) Stop, if the number of selected probes surmounts a given threshold or if each partition contains only one sequence. The latter case can be called *complete partitioning* of the training set.

The criterion to assess the partitioning produced by a set of probes is the entropy of the partitions defined above. It attains a maximum, $I = \log_2 N$, if all sequences have different fingerprints.

The algorithm takes as input the training set, i.e. a number of sequences, for example in the FASTA-format. The number of training sequences can vary but training should not be done with less than 500 sequences. This number is a heuristic threshold which ensures that all octamer probes have a sufficient chance to occur in the training set and that differences in probe occurrences can be identified. The threshold might be set lower in some applications especially when too few sequences are available or if the probe lengths are smaller than 8 bp. If the training set is too small it is likely that useful probes are excluded from further analysis simply because of their absence in the training set. The algorithm runs with several parameters that can be specified by the user:

- LEN: length of probes. Default LEN = 8.
- N_GC: minimal number of G + C in each probe. Default N_GC = 2.
- COMP: minimal complexity of probes (see explanation below). Default COMP = 0.5.
- OVL: maximal length of common stretch of basepairs shared by any two probes. Default 0.75*LEN, i.e. OVL = 6.

Table 2. First 12 probes selected from 6000 human cDNA training sequences, the number of different fingerprints achieved with these probes and the total entropy of the resulting partition

Number of probes	Probe sequence	Different fingerprints	Entropy
1	CAGTAATA	2	0.783558
2	CAGCCTGG	4	1.534318
3	CCAGCCCC	8	2.234960
4	CTGGGGCC	16	2.914776
5	AGCAGCAG	31	3.581627
6	CAGCTCCA	57	4.225269
7	CAGCCTCC	106	4.839472
8	CCTGCAGC	192	5.439841
9	CCCTGGCC	339	6.017425
10	CCCTGGAG	542	6.563812
11	GATGGTGA	745	7.067720
12	CCAGCTGC	1028	7.530276
13	—	—	—

- SEL: number of probes to be determined. Default SEL = 200.

To enhance the practical use of the algorithm we allow the user to exclude probes that match with undesirable sequences. For example if parts of the vector sequence of the clones are amplified by PCR then a probe that matches with the amplified part of the vector will hybridise in practical experiments to nearly all cDNA clones which can disturb further analysis. The user can thus specify a file that contains undesirable sequences. Sequences should be in the same format as the training sequences. Probes that match with any of these undesirable sequences are excluded from further analysis. Additionally it has been shown in the context of designing PCR primer pairs that it is useful if hyper-abundant RNAs such as ribosomal RNAs, mitochondrial RNAs and dispersed repeats like SINEs or LINEs are excluded from analysis (Pesole *et al.*, 1998).

The output of the program is a list of probes (appearing in the order of the selection process), the number of different fingerprints achieved and the total entropy of the corresponding partition. Table 2 gives an example of the first twelve probes selected from 6000 human training sequences.

It can be observed that the actual number of different fingerprints achieved is far less than the number of theoretically possible fingerprints. Only 1028 out of the $2^{12} = 4096$ different fingerprints occur within the set of sequences (~25%). This rate drops rapidly since partitions represented by fingerprints with a large number of positives are very unlikely to occur due to the low matching frequencies of octamer probes (see calculation below).

The runtime of the algorithm on a Digital-alpha 500 MHz computer has been 154 minutes and 74 minutes for designing 200 probes and 100 probes respectively from 6000 sequences and six minutes for selecting 30 probes from 500 sequences. Memory increases with the length of the probes, because the sequence for each possible oligo is stored, and with the number of training sequences. After pre-selecting probes according to G + C-content and complexity the 2000 most frequent probes according to the training set are kept for the selection process. This number is heuristic and can be modified by the user, but it is reasonable to restrict the number of admissible probes because it is computational unfeasible to process all possible probes especially when the probe lengths increase (for example there are 32 768 different 8-mers but 524 288 different 10-mers when sequences and reverse complementary sequences are counted only once). On the other hand the information gain of low-frequent probes is decreasing rapidly so that these probes are not relevant for the selection process. Memory scales, thus, mainly with the number of training sequences. It has been 30 MB for designing probes from 6000 sequences and 14 MB for designing probes from 500 sequences.

Data sets

A total number of (a) 12 000 human and (b) 12 000 rodent cDNA sequences were chosen randomly from the GenBank database. Only sequences longer than 300 bp were chosen, sequences longer than 2000 bp were cut at that level so that the actual range of sequence length is the interval [300, 2000]. This selection is due to our experimental observations where we find PCR-products of cDNA clones within this range. Sequences were subjected to a cleaning procedure using the program CLEANUP (Grillo *et al.*, 1996) with default parameters in order to remove homologous sequences. A subset of 6000 sequences was used as a training set in both (a) and (b) for designing the different probe sets and additionally five different test sets from sequences non-overlapping with the training sequences are created to test cluster performance. We introduce a cluster structure as described above so that each of the five human and rodent data sets has the size of 2099 copies created from 685 different cDNA's. We calculate binary fingerprints according to 200 octamer probes from probe sets P_{ent}, where probes are derived by maximising the entropy, P_{freq}, where the most frequent probes are chosen and P_{rand}, where a random collection of probes is used. We excluded probes with low complexity (COMP < 0.5, see explanation below) such as TATATATA and probes with a G + C-content of less than two in order to focus selection on probes that are of practical use in our hybridisation experiments.

Each of the five human (and rodent) data sets is clustered with each of the probe sets 30 times to derive variations in clustering quality; variation is due to the fact that noise is spread randomly for each probe via false positive and false negative rate in each simulation run. Our simulation results are thus based on a total of 900 simulation runs.

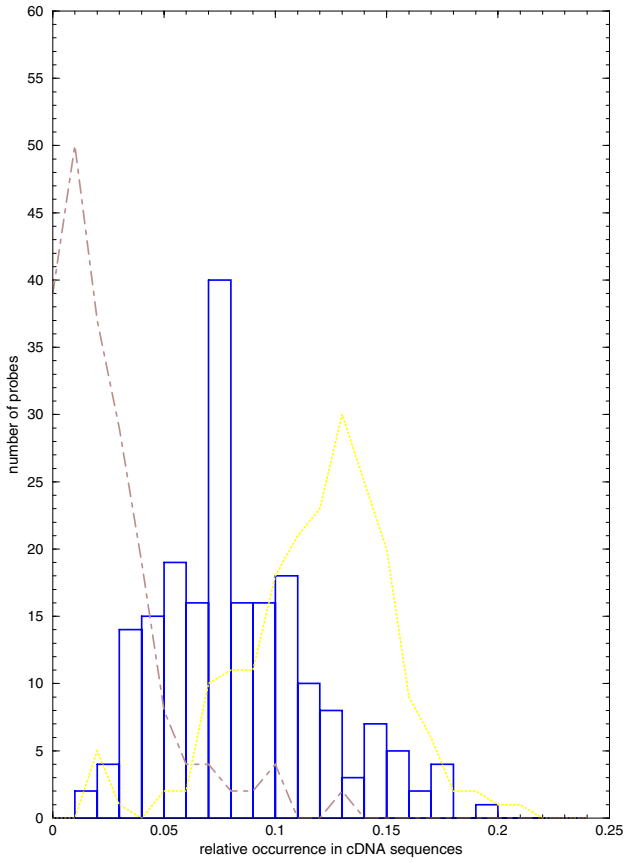

Fig. 2. Frequency distribution of probe sets within ~7000 human cDNA sequences of length 1000 bp. The histogram (solid) shows the relative occurrences of 200 probes selected by entropy maximisation, the dotted line shows the relative occurrences of 200 probes selected by frequencies and the dot-dashed line shows the distribution for 200 randomly chosen probes.

Results

Frequency of probes

Figure 2 shows a histogram on the occurrence of the respective probe sets in 7000 human sequences of equal lengths 1000 bp.

We observe that both P_{ent} and P_{freq} show a significant increase in matching frequency as compared to the randomly chosen set P_{rand}. We can compare our practical results with theoretical considerations: given N sequences of equal lengths L and a probe of length l, the expected frequency of the probe can be modeled using a binomial

distribution. If we take into account that it is not relevant if the probe or its reverse complementary sequence hits the clone sequence and that all clone sequences are different from each other then each of the $M \sim \frac{4^l}{2}$ probes has the same probability of occurring in each of the sequences. The number of matches of a probe within the N sequences can then be viewed as the outcome of N independent and identically distributed *Bernoulli*-experiments $X_1, \ldots, X_N$ with probability of success, p, equal to

$$p = \text{prob}\,(X_i = 1) = 1 - \left(\frac{M-1}{M}\right)^{L-l+1}.$$

The probability of success for matching an individual sequence is derived by the following consideration: the probe can match in any of the $L - l + 1$ positions of the sequence. It does not match with the sequence if all positions are held by any of the $M - 1$ other possible probes. The probability of that event is equal to $\left(\frac{M-1}{M}\right)^{L-l+1}$, the probability of the complementary event thus yields the result for p. The expected number of successes within the sequences is then given by $E\left(\sum_i X_i\right) = N^* p$. For example, if we work with octamers ($l = 8$) and we have sequence lengths $L = 1000$ bp the expectation is $E\left(\sum_i X_i\right) \sim N^*0.030$, which means that a randomly chosen octamer probe will approximately hybridise to 3% of the sequences if they are independent of each other. This calculation corresponds to the dot-dashed curve in Figure 2. This matching rate is too low to allow good discrimination of sequences by fingerprints so that strategies are necessary that increase matching frequency. Frequency can of course be increased by reducing l, the length of probes. For example the expected frequency of heptamer probes is $E\left(\sum_i X_i\right) \sim N^*0.11$, so that a randomly chosen heptamer probe will approximately match to 11% of the sequences. But the hybridisation stability of most heptamer probes is not sufficient in practice. Figure 2 shows that both sets, P_{ent} and P_{freq}, contain octamer probes with a matching frequency in the order of randomly chosen heptamer probes.

Comparison of probe sets

Figure 3 shows the results of the simulation procedure when comparing the different probe sets (human sequences 3a, rodent sequences 3b). It is obvious that in both cases P_{ent} (solid lines) is superior to P_{freq} (dotted lines) and P_{rand} (dot-dashed lines), where P_{freq} is also significantly better than P_{rand}. For each data set 30 different clustering runs were performed to derive the mean, m, and the standard deviation, s, of the quality measure. The bars indicate the interval $[m - s, m + s]$. Furthermore it can be observed that variation of quality within each test set is decreased when using P_{ent} instead of P_{freq}; e.g. test set 2 of the human data shows high variation with P_{freq} but not with P_{ent} . P_{rand} performs poorly on most data sets compared to the other probe sets except on set 5 of the rodent data; here it is remarkable that clustering quality is even better. Further investigation showed that there are a lot of short sequences within the data set (<400 bp) so that we think this result is rather an artefact of short lengths of the sequences in this test set.

Figure 3c shows furthermore that the entropy maximised probe set, $P_{\text{ent}}^{\text{hum}}$, of the human training set (solid line) behaves better than the respective probe set, $P_{\text{ent}}^{\text{rod}}$, derived from rodent sequences (dotted line) when applied to the human test sets. We could not expect variation to be very significant because of the phylogenetic neighbourhood of human and rodents but still it is marked. Differences are more significant when organisms are used that have a greater phylogenetic distance. We add therefore results for 200 probes, $P_{\text{ent}}^{\text{plant}}$(dot-dashed line), that are trained on 40 000 plant EST's extracted from *GenBank* (core eudicots); in this case we observe a more significant decrease in performance especially in test sets 1, 2 and 5. Quality of $P_{\text{ent}}^{\text{hum}}$ and $P_{\text{ent}}^{\text{rod}}$ is comparable within four of the five test sets but it drops significantly in test set 2 (<0.7); in this case the use of plant-trained probes is inefficient. We thus conclude that for each organism under analysis probe sets should be calculated from training sets matched as closely as possible to achieve the best performance of the fingerprinting procedure.

Variation of algorithmic parameters

Figure 4 shows the performance of the calculated probe set in generating different fingerprints in the human training sequences when varying internal algorithmic parameters. Because of runtime performance we restricted the number of training sequences to 500 sequences.

The complexity of a probe is calculated according to the entropy of the dimer composition of the probes. From an oligomer of length l we can extract $l - 1$ dimers, i.e. the maximal complexity of an $l$mer probe is equal to $\log_2(l - 1)$. The complexity is normalised so that its range is between 0 and 1. The octamers TTGACTAA, TACGACAC and TATATATA, for example, have complexities 1.0, 0.7580 and 0.3509 respectively. Figure 4a shows the performance if probes are pre-selected according to complexity, i.e. only probes with complexities $\geqslant$ COMP are admissible. A complete partitioning of the 500 sequences can be performed with 27 probes setting COMP = 0.0 (pluses) and with 30 probes setting COMP = 0.75 (diamonds) so that an additional amount of ~10% of probes is necessary when setting the threshold so that most repeat sequences such as TATATATA are excluded (COMP = 0.75). This is not very restrictive

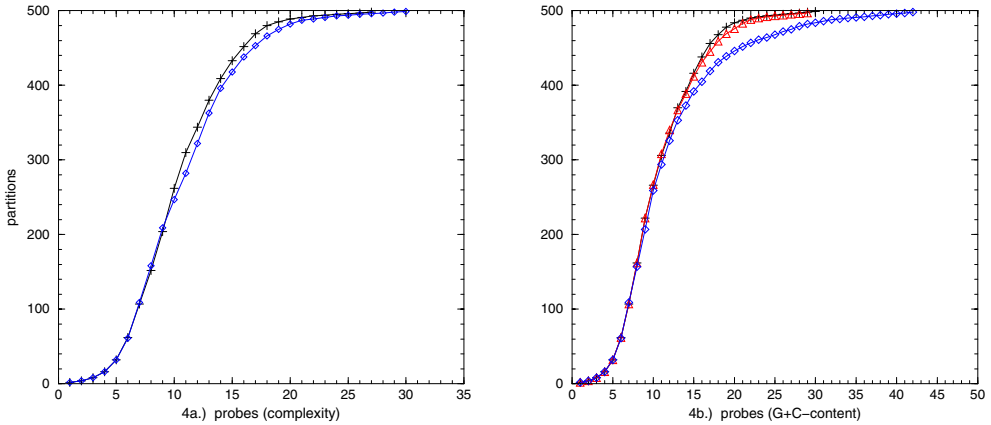

Fig. 4. Variation of algorithmic parameters. (**a**) Partitioning performance of successively selected probes within 500 human cDNA sequences when probes are pre-selected after dimer-complexity (COMP = 0.0 (pluses) and 0.75 (diamonds)). y-axis shows the number of different fingerprints achieved. (**b**) Partitioning performance when probes are pre-selected after G + C-content (N_GC = 2 (pluses), 4 (triangles) and 6 (diamonds)).

so that the default value for minimal complexity of the probes can be set quite generously.

It is experimentally crucial that probes have a good hybridisation stability. This is usually assured by selecting probes according to G + C-content because guanine and cytosine form more stable base pairs. Clearly this constraint restricts the number of possible probes. Figure 4b shows how the number of different fingerprints decreases if the minimal G + C-content of octamers, N_GC , is set to 2 (pluses), 4 (triangles) and 6 (diamonds). A complete partitioning has been achieved with 27 probes using N_GC = 2, whereas for N_GC = 6 42 probes are necessary. This parameter affects partitioning to a higher degree so that pre-selection of probes according to G + C-content should be done with a moderate threshold.

Discussion and conclusion

We have shown that the choice of probes is crucial for the identification of clone sequences by hybridisation experiments because it influences the quality of clustering of the resulting fingerprints to a high degree. Our method performs substantially better than the set of most frequent probes and randomly chosen probes. This is neither an artefact of the individual test set as is shown by repeating clustering on different test sets nor is it an artefact of the training sets as is shown by testing sequences from different organisms.

The value of the method lies in the fact that it evaluates the probe set as a whole not only the individual probe sequences. For example if the probe sequence AAGCAGTT has high matching frequency then it is very likely that the probe sequence TTGCAGTT has high matching frequency too. Both sequences would have been chosen as the most frequent probes. The gain in information however is not very high so that the total entropy of the partitions induced by both probes would not be high.

Differences in clustering performance will further increase with the overall matching frequencies of the probes. A possibility of increasing matching frequencies is the use of shorter probes such as 7-mers. Unfortunately experiments with radioactively labeled heptamer DNA-probes showed poor hybridisation stability. Technical developments are therefore ongoing in many laboratories to develop shorter DNA-analogues like PNA's or LNA's (Egholm *et al.*, 1993; Singh *et al.*, 1998). Achieving higher hybridisation rates will support our method of selecting probes according to their partitioning quality.

It has also been shown that the information theoretical partitioning of clone sequences is dependent on the training set, i.e. the training set should be chosen as close to the organism under analysis as possible. We plan to produce probe lists for several organisms (e.g. human, mouse, sea-urchin, zebra-fish) that are highly informative for the respective sequences; when starting hybridisation experiments with a special tissue it is recommended to use the designed probe sets. Besides screening cDNA data from specific tissues the algorithm can be used to train probes on genomic data. An application to human shot-gun libraries was published recently (Radelof *et al.*, 1998). The construction of the training set in the case of genomic data depends on the application. In this work for example we retrieved some megabases of publicly available DNA, cut it into several thousands of pieces with size close to the insert sizes of the experimental material (∼1200 bp) and used those simulated 'shot-gun' sequences as a training set.

As our results are produced solely by simulations we can not guarantee that all probes have a sufficient hybridisation

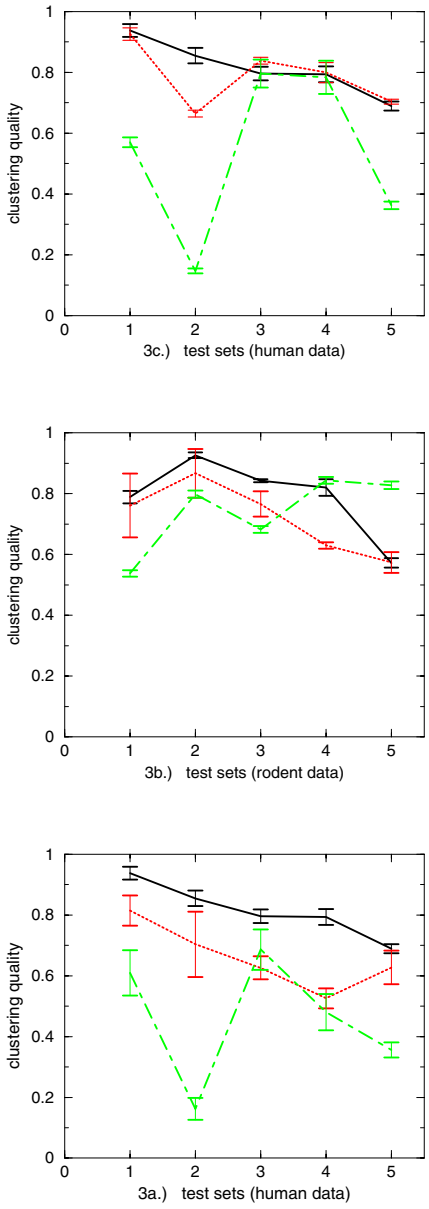

Fig. 3. Clustering performance of probe sets. (**a**) five different human cDNA test sets are clustered with P_{ent}, the set of 200 entropy-maximised probes (solid line), P_{freq}, the 200 most frequent probes (dotted line), and P_{rand}, 200 randomly chosen probes (dot-dashed line); probe sets are derived from the human training set. Clustering quality is calculated using the Jaccard-coefficient which is high if clustering quality is good (1 = perfect clustering). For each test set 30 different clustering runs are performed to derive the mean, m, and the standard deviation, s, of the quality indices. Bars show the interval $[m-s, m+s]$. (**b**) Five different rodent cDNA test sets clustered with P_{ent} (solid line), P_{freq} (dotted line) and P_{rand} (dot-dashed line) derived from the rodent training set. (**c**) Performance between organisms; P_{ent}^{hum} derived from the human training data (solid line) versus P_{ent}^{rod} derived from the rodent training data (dotted line) and P_{ent}^{plant} derived from plant training data (dot-dashed line) tested on human test sets.

quality in practice. Good hybridisation quality can be achieved, e.g. by G + C-rich probes. We have shown how the G + C-content influences the partitioning quality of the probes so that the number of probes should be increased to compensate this effect. A certain amount of probes will fail to give good results in practice due to unfavorable hybridisation conditions. Our experiments suggest a rate of 15–20%. It is therefore convenient to add a common stock of probes that yield good practical results to the tissue-specific probes. In some situations it might also be desirable to add to the probe list special probes e.g. motif-oligos that recognize specific words in the sequences under analysis.

Validation tools presented in this work are mainly derived from our simulation pipeline because our main interest is the influence of the probe sets on clustering and clone identification. The probe set is one parameter of the simulation pipeline and when fixing all other parameters we can extract information about performance of probe sets. We have shown only a small part of the simulation pipeline which has a number of different parameters describing the probe–target interaction (e.g. mismatches, clone concentration, degree of radioactive labeling, scanning procedure). Work is ongoing to measure the influence of these parameters on the resulting clustering quality (unpublished).

The main application shown in this paper has been oligo-fingerprinting but the algorithm can be applied to any experiment where one wishes to identify sequences by short probes via the occurrence of those probes in the sequences or—more generally—experiments where texts (sequences) are characterised by words (probes); a possible application will be database searches for text documents by words, where the algorithm might be used to select characteristic keywords from a suitable training set.

Acknowledgements

The authors thank M.Clark, G.Panopoulou and A.Poustka for many discussions on hybridisation topics. This work has been financed by the Max-Planck Society.

References

Clark,M.D., Panopoulou,G.D., Cahill,D.J., Büssow,K. and Lehrach,H. (1999) Construction and analysis of arrayed cDNA libraries. In Weissman,S.M. (ed.), *Methods in Enzymology* vol. 303, Academic Press, San Diego, pp. 205–233.

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. Wiley, New York.

Cuticchia,A.J., Arnold,J. and Timberlake,W.E. (1993) PCAP: probe choice and analysis package—a set of programs to aid in choosing synthetic oligomers for contig mapping. *CABIOS*, **9**, 201–203.

Drmanac,S., Stavropoulos,N.A., Labat,I., Vonau,J., Hauser,B.,

Soares,M.B. and Drmanac,R. (1996) Gene-representation cDNA clusters defined by hybridization of 57 419 clones from infant brain libraries with short oligonucleotide probes. *Genomics*, **37**, 29–40.

Egholm,M., Buchardt,O., Christensen,L., Behrens,C., Freier,S.M., Driver,D.A., Berg,R.H., Kim,S.K., Norden,B. and Nielsen,P.E. (1993) PNA hybridizes to complementary oligonucleotides obeying the Watson-Crick hydrogen-bonding rules. *Nature*, **365**, 566–568.

Fu,Y.X., Timberlake,W.E. and Arnold,J. (1992) On the design of genome mapping experiments using short synthetic oligonucleotides. *Biometrics*, **48**, 337–359.

Grillo,G., Attimonelli,M., Liuni,S. and Pesole,G. (1996) CLEANUP: a fast computer program for removing redundancies from nucleotide sequence database. *CABIOS*, **12**, 1–8.

Herwig,R., Poustka,A.J., Müller,C., Bull,C., Lehrach,H. and O'Brien,J. (1999) Large-scale clustering of cDNA fingerprinting data. *Genome Res.*, **9**, 1093–1105.

Lehrach,H., Drmanac,R., Hoheisel,J., Larin,Z., Lennon,G., Monaco,M.P., Nizetic,D., Zehetner,G. and Poustka,A. (1990) Hybridization fingerprinting in genome mapping and sequencing. In Davies,K.E. and Tilghman,S. (eds), *Genome Analysis Volume 1: Genetic and Physical Mapping* Cold Spring Laboratory Press, Cold Spring Harbor, New York, pp. 39–81.

Lennon,G. and Lehrach,H. (1991) Hybridization analyses of arrayed cDNA libraries. *Trends Genet.*, **7**, 314–317.

Maier,E., Meier-Ewert,S., Ahmadi,A., Curtis,J. and Lehrach,H. (1994) Application of robotic technology to automated sequence fingerprint analysis by oligonucleotide hybridisations. *J. Biotechnol.*, **35**, 191–203.

Meier-Ewert,S. (1994) Global expression mapping of mammalian genomes, *PhDThesis*, University College, London.

Meier-Ewert,S., Maier,E., Ahmadi,A., Curtis,J. and Lehrach,H. (1993) An automated approach to generating expressed sequence catalogues. *Nature*, **361**, 375–376.

Meier-Ewert,S., Lange,J., Gerst,H., Herwig,R., Schmitt,A.O., Freund,J., Elge,T., Mott,R., Herrmann,B. and Lehrach,H. (1998) Comparative gene expression profiling by oligonucleotide fingerprinting. *Nucleic Acids Res.*, **26**, 2216–2223.

Mirkin,B. (1996) *Mathematical Classification and Clustering*. Kluwer AcademicPublishing, Dordrecht.

Milosavljevic,A., Zeremski,M., Strezoska,Z., Grujic,D., Dyanov,D., Batus,S., Salbego,D., Paunesku,T., Soares,M.B. and Crkvenjakov,R. (1996) Discovering distinct genes represented in 29 570 clones from infant brain cDNA libraries by applying sequencing by hybridization methodology. *Genome Res.*, **6**, 132–141.

Panopoulou,G.D., Clark,M.D., Gerst,H., Herwig,R., Holland,L.Z., Holland,N.D. and Lehrach,H. (1998) Large-scale identification of amphioxus genes from different developmental stages using oligonucleotide fingerprinting. *Develop. Biol.*, **198**, 200–201.

Pesole,G., Liuni,S., Grillo,G., Belichard,P., Trenkle,T., Welsh,J. and McClelland,M. (1998) GeneUp: a program to select short PCR primer pairs that occur in multiple members of sequence lists. *BioTechniques*, **25**, 112–123.

Poustka,A., Pohl,T., Barlow,D.P., Zehetner,G., Craig,A., Michiels,F., Ehrich,E., Frischauf,A.M. and Lehrach,H. (1989) Molecular approaches to mammalian genetics. In *Cold Spring Harbor Symposia on Quant. Biol.* vol. 51, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 131–139.

Poustka,A.J., Herwig,R., Krause,A., Hennig,S., Meier-Ewert,S. and Lehrach,H. (1999) Towards the gene catalogue of sea urchin development: the construction and analysis of an unfertilized egg cDNA library highly normalized by oligonucleotide fingerprinting. *Genomics*, **59**, 122–133.

Radelof,U., Hennig,S., Seranski,P., Steinfath,M., Ramser,J., Reinhardt,R., Poustka,A., Francis,F. and Lehrach,H. (1998) Preselection of shotgun clones by oligonucleotide fingerprinting: an efficient and high throughput strategy to reduce redundancy in large-scale sequencing projects. *Nucleic Acids Res.*, **26**, 5358–5364.

Shannon,C. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.

Schmitt,A.O., Herwig,R., Meier-Ewert,S. and Lehrach,H. (1999) High-density cDNA grids for hybridization fingerprinting experiments. In Innis,M.A., Gelfand,D.H. and Sninsky,J.J. (eds), *PCR Applications. Protocols for Functional Genomics* Academic Press, San Diego, pp. 457–472.

Singh,S.K., Nielsen,P., Koshkin,A.A. and Wengel,J. (1998) LNA (Locked Nucleic Acids): synthesis and high-affinity nucleic acid recognition. *Chem. Commun.*, 455–456.