BIOINFORMATICS

Vol. 22 no. 14 2006, pages e424–e430
doi:10.1093/bioinformatics/btl254

Decoding non-unique oligonucleotide hybridization experiments of targets related by a phylogenetic tree

Alexander Schliep[1,*] and Sven Rahmann[2,3,*]

[1]Dept. Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63–73, D-14195 Berlin, Germany, [2]Algorithms and Statistics for Systems Biology group, Genome Informatics, Technische Fakultät, Universität Bielefeld, D-33594 Bielefeld, Germany and [3]International NRW Graduate School of Bioinformatics and Genome Research, Universität Bielefeld

ABSTRACT

Motivation: The reliable identification of presence or absence of biological agents ("targets"), such as viruses or bacteria, is crucial for many applications from health care to biodiversity. If genomic sequences of targets are known, hybridization reactions between oligonucleotide probes and targets performed on suitable DNA microarrays will allow to infer presence or absence from the observed pattern of hybridization. Targets, for example all known strains of HIV, are often closely related and finding *unique* probes becomes impossible. The use of *non-unique* oligonucleotides with more advanced decoding techniques from statistical group testing allows to detect *known* targets with great success. Of great relevance, however, is the problem of identifying the presence of previously *unknown* targets or of targets that evolve rapidly.

Results: We present the first approach to decode hybridization experiments using non-unique probes when targets are related by a phylogenetic tree. Using a Bayesian framework and a Markov chain Monte Carlo approach we are able to identify over 94% of known targets and assign up to 70% of unknown targets to their correct clade in hybridization simulations on biological and simulated data.

Availability: Software implementing the method described in this paper and datasets are available from http://algorithmics.molgen.mpg.de/probetrees.

Contact: alexander.schliep@molgen.mpg.de, Sven.Rahmann@cebitec.uni-bielefeld.de

1 INTRODUCTION

Identifying biological targets. Identifying viruses infecting a patient, detecting bacteria spoiling food, or deciding whether a water sample is safe for humans to drink are tasks which share the same underlying problem: to identify certain *targets* in biological (DNA) samples. Targets refer to the biological agents, the viruses, bacteria or other organisms that we want to detect. Recent developments in the Avian influenza pandemic brought virus identification into the front-news spotlight. In addition to accurately determining the lethal virus strain [Putonti *et al*., 2006], it is crucial to screen humans and animals, which might host several viruses and thus allow cross-species recombination. More optimistic applications of target detection are the study of biodiversity, say on the microbial level, and environmental microbiology. The target identification problem is also central in the area of biothreat reduction.

In clinical applications, target identification has classically been achieved for individual targets with unique markers such as staining techniques for specific antibodies. While one test per potential target is acceptable for many medical applications, it is not a cost-effective strategy if the number of potential targets is large, if several targets might be present simultaneously, or if many samples must be investigated. In South Africa for example, HIV super-infections, i.e., simultaneous infections with multiple HIV strains, are much more prevalent than in the Western world. In these cases, clinical marker kits for strain identification are more prone to failure.

Approaches based on unique probes. One experimental assay widely used in molecular biology is the hybridization reaction of fluorescently labeled DNA or RNA molecules to complementary DNA or RNA. Such hybridization reactions can be used for target detection if (partial) genomic sequences of targets are available. Often, short oligonucleotide DNA microarrays are used as technology platform (the approach in principle generalizes to other hybridization-based technologies). Assuming ideal conditions, we would select one specific oligonucleotide probe that hybridizes to its intended target only and does not cross-hybridize to any other target. Subsequently, we detect presence and absence of targets in a sample from the observed hybridization pattern. This *unique probe* approach has been originally developed for the design of gene expression DNA microarrays using oligonucleotide probes (e.g., [Kaderali and Schliep, 2002; Rahman, 2003a]). However, in the applications described above, targets are often closely related and thus unique probes cannot be found.

Non-unique probes. The use of *non-unique probes*, hybridizing to several targets simultaneously, poses problems in the analysis of experiments. If one assumes that at most one target can be present simultaneously, the problem can be handled effectively [Wang *et al*., 2003, Rash and Gusfield, 2002]. This assumption is unrealistic, however, and [Schliep *et al*., 2003] introduced a *statistical group-testing* approach to address the case when multiple targets are present simultaneously. Subsequent work [Klau *et al*., 2004] has attempted to minimize the number of probes required to reliably identify small-cardinality target sets by an integer linear programming approach. In all of the above work, only the ability to detect *known* targets has been evaluated.

[*]To whom correspondence should be addressed.
Both authors contributed equally.

Novel contributions. We extend the group-testing approach using non-unique probes to targets related by a phylogenetic tree. This allows us to consider an intriguing and highly relevant question: Can we even detect the presence of yet *unknown* targets, e.g., can we detect the presence of a new strain, or can we detect the presence of a known target if it (and its hybridization pattern) has changed because of fast evolution? Even if we restrict ourselves to a specific virus, the targets used as input will only represent a sample of all existing strains and new strains are likely to arise between the time of microarray design and its large-scale use. To our knowledge, this article is the first work to address these issues.

Outline. We describe the probe selection strategy and group testing methods in Section 2, particularly focusing on the novel aspect how they can be integrated with phylogenetic tree information. Section 3 presents artificial and real datasets for evaluating these methods, describes our evaluation criteria, and shows the evaluation results. A concluding discussion is given in Section 4.

2 METHODS AND MODELS

Notational Remarks. If $\mathcal{S}$ is a finite set, $|\mathcal{S}|$ is its cardinality. We identify binary vectors $T \in \{0,1\}^m$ with the index set $\{i : T_i = 1\} \subset \{1,\ldots,m\}$, for which we also write T, so $|T| = \sum_{i=1}^{m} T_i$.

2.1 Overview: Problem setting

Initially, we are given a set $\mathcal{T}$ of DNA target sequences (the *known targets*) and a phylogenetic tree $\mathcal{B}$ relating them. Depending on the application, the targets might be whole genomes (e.g., all known HIV strain genomes), or single gene sequences (e.g., the cytochrome C sequences of several related species). We assume that the target set contains many closely related and hence similar sequences.

Our objective is to be able to decide which of these targets are present and which ones are absent in unclassified DNA samples when we observe an oligonucleotide probe hybridization fingerprint for the sample. To be more precise, we assume that we observe which probes react to some target(s) in the sample, but that this observation is *noisy*. In most applications, we may assume that the target set contained in the sample is small compared to the whole set $\mathcal{T}$ (e.g., the set of HIV strains infecting a single patient).

Additionally, we expect that the sample may contain *unknown targets*, that is, sequences similar to those in $\mathcal{T}$ that were not available when $\mathcal{T}$ was prepared. This would be the case for new virus strains or fast evolving genomes, for example. Although we cannot expect to perfectly classify these unknown targets, we would at least like to place them at the correct location in the tree $\mathcal{B}$.

Our first tasks are thus

(1) to select suitable *probe candidates* for the given target set $\mathcal{T}$. Note that the usual probe design methodologies that look for target-specific probes do not have a good chance of success on the typical datasets we consider: Because of the high sequence similarity between targets, only very few specific probes will be found. Our proposed solution is to use a *group testing* approach that allows *non-unique* probes. We deal with the ensuing complications in a subsequent *decoding* step. The candidate selection step also ensures that no probes are selected that could hybridize to genomes of contaminating organisms or host organisms (e.g., the human genome for HIV viruses);

(2) to reduce the candidate set to a final *probe set* $\mathcal{P}$;

(3) to compute the $|\mathcal{T}| \times |\mathcal{P}|$ *basic hybridization matrix* H^{basic}, a binary matrix defined by $H_{ij}^{\text{basic}} = 1$ if target i hybridizes to probe j, and $H_{ij}^{\text{basic}} = 0$ otherwise;

(4) to extend the hybridization patterns (rows) of H^{basic} from targets to whole subtrees (monophyletic groups) of $\mathcal{B}$ by deciding which hybridization pattern would be ‘‘typical’’ for unknown targets in a monophyletic group. We obtain an *(extended) hybridization matrix* H of size $(|\mathcal{T}| + |\mathcal{I}|) \times |\mathcal{P}|$, where $\mathcal{I}$ denotes the set of internal (non-leaf) nodes of $\mathcal{B}$.

The above steps are described formally in Sections 2.2 (probe selection) and 2.3 (computing H), followed by a small example.

Given H and a target set $T \subset \{1,\ldots,|\mathcal{T}| + |\mathcal{I}|\}$, it is straightforward to compute the theoretical (i.e., error-free) hybridization result $r = r(T) \in \{0,1\}^{|\mathcal{P}|}$: We will observe $r_j = 1$ if there exists a target $i \in T$ to which probe j hybridizes ($H_{ij} = 1$). In other words, $r_j = \vee_{i\in T} H_{ij}$, so r is the logical or of the rows indicated by T. In reality, however, we need to take noisy results into account: Probes not showing a hybridization signal although they should are called *false negatives*, and probes showing a signal although they should not are called *false positives*. The error model is described in Section 2.4.

For an unidentified DNA sample, we need to solve the inverse problem of the above one: We observe a certain result r, and our task is to find T, which may consist of both known targets $t \in \mathcal{T}$ and unknown targets $t \in \mathcal{I}$ modeled by internal nodes of $\mathcal{B}$, such that T best explains r. We adopt a Bayesian framework and introduce a target set prior in Section 2.5. Then our goal becomes to find the target set that maximizes the posterior probability given r, which turns out to be a difficult problem to solve exactly. We thus switch to a Gibbs sampling strategy, which we describe in Section 2.6.

2.2 Probe selection

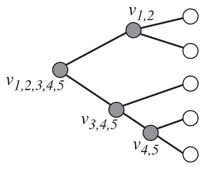
We start with a set $\mathcal{T} = \{t_1,\ldots,t_m\}$ of m distinct but similar DNA sequences, the *targets*. The first step is to find characteristic substrings (the *probes*) either for single targets or for whole target sets $T \subset \mathcal{T}$. The idea is that an unidentified DNA sample can be tested quickly and (relatively) cheaply for the occurrence of all probe sequences, e.g., by a microarray hybridization experiment, whereas determining the precise sequences of all sample members would be a more complicated procedure.

A good (specific or unique) probe p is characterized by the fact that it hybridizes well to a single target and not at all to the remaining targets. Because of the high sequence similarity in $\mathcal{T}$, however, unique probes will be difficult to find in sufficient number. Instead of attempting a bad compromise, we turn this problem into a feature and allow that p hybridizes to a small group $\mathcal{T}_p$ of targets; this need not be a monophyletic group in $\mathcal{B}$. We require, however, that the probe makes a clear distinction between $\mathcal{T}_p$ and $\mathcal{T} \setminus \mathcal{T}_p$ in the sense that there is a strong observable signal for all $t \in \mathcal{T}_p$ and no signal for all $t \in \mathcal{T}\setminus\mathcal{T}_p$.

The dynamics of DNA-DNA hybridization are quite complicated and not fully understood. However, it is reasonable to assume that a probe will give a clear positive signal if it is an exact substring of the target, and that no signal will be observed if the longest common substring between probe and target is very short. This so-called *longest common factor approach* was first proposed in [Rahmann, 2003a, 2002] and provides a practical and efficient surrogate measure for the true probe-target affinity. What must be avoided are probes that have long but not full-length common substrings with some targets in $\mathcal{T}$.

We thus proceed as follows. Every substring p in a given length range (our method is mainly applicable to short oligonucleotides between 20 and 30 nt) of any target in $\mathcal{T}$ is tested against the other targets for long (but not full-length) common substrings and discarded as a probe candidate if any are found. For the remainder, the hybridization stability (Gibbs free energy) is estimated using the nearest-neighbor model described in [SanataLucia, 1998]. The probes are accepted only if their estimated Gibbs free energy falls into a small homogeneous range to ensure similar hybridization behavior. All of these steps are implemented in the existing PROMIDE software described in [Rahmann, 2003a]. The main reason to choose PROMIDE is that it is one of the few programs that allows non-uniqe probe selection.

The nature of the selection process allows to model hybridization as a yes/no event that can be described by a binary matrix H^{basic}: Consider the



	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
t_1	0	1	0	1	0	0	1	0	1	1
t_2	0	0	1	1	1	1	1	0	1	0
t_3	1	1	0	1	1	1	0	0	1	0
t_4	1	1	1	1	0	1	0	1	0	1
t_5	1	1	1	1	0	1	0	1	1	0
$v_{1,2}$	0	0	0	1	0	0	1	0	1	0
$v_{1,2,3,4,5}$	1	1	1	1	0	1	0	0	1	0
$v_{3,4,5}$	1	1	1	1	0	1	0	1	1	0
$v_{4,5}$	1	1	1	1	0	1	0	1	0	0

	T	Hybridization result r
a.	$\{1\}$	0 1 0 1 0 0 1 0 1 1
b.	$\{1,2\}$	0 1 1 1 1 1 1 0 1 1
c.	$\{v_{1,2}\}$	0 0 0 1 0 0 1 0 1 0
d.	$\{v_{1,2},3\}$	1 1 0 1 1 1 1 0 1 0
e.	$\{2,4\}$	1 1 1 1 1 1 1 1 1 1
f.	$\{1,2,5\}$	1 1 1 1 1 1 1 1 1 1
g.	???	1 0 1 0 1 0 1 1 0 1

Fig. 1. *Left*: A small hybridization matrix H. Rows 1–5 define a hypothetical *basic hybridization matrix* H^{basic}, as it would result from a probe selection process. Rows 6–9 are associated to the internal nodes of the phylogenetic tree $\mathcal{B}$ shown to the left of H^{basic}. They are computed as strict majority functions and represent any so far unknown target that could exist in the monophyletic group below the respective node. *Right*: Seven examples (a–g) of sets of known and unknown targets and their expected hybridization results (the or of the rows indicated by the target set); see Section 2.4 for details.

relation of target i and probe candidate j: Either the probe is a substring of the i (in which case we assume a stable hybridization and set $H_{ij}^{\text{basic}} := 1$), or they share only a short common substring (in which case stable hybridization does not occur and we set $H_{ij}^{\text{basic}} := 0$). The intermediate case where ‘‘almost’’ the whole probe occurs in some target is ruled out by the longest-common-factor-based selection process.

In the resulting set of probe candidates, maybe no probe identifies any target uniquely, but certain combinations of probes still identify certain combinations of targets (target sets). This places our approach in the field of *group testing*: Each probe tests whether *some* member of a certain target group is present, but cannot tell *which one*. The resulting decoding problem is described in Section 2.6.

Reducing the set of probe candidates. In previous work [Klau *et al.*, 2004], we have shown that the resulting probe candidate set can often be reduced considerably (up to 50%) without sacrificing decoding resolution if the probes are picked carefully. However, in this study, we face a different task: One of our goals is to detect unknown or fast-evolving targets. Therefore, any optimization of the probe set, even if it does not adversely affect our ability to identify known targets, would certainly decrease our chances of identifying unknown targets. Therefore, we have not reduced the probe set.

2.3 Extending the hybridization matrix to monophyletic groups

After probe selection, we have n probes $\mathcal{P} = \{p_1, \ldots, p_n\}$ for the m targets $\mathcal{T} = \{t_1, \ldots, t_m\}$, we know the basic hybridization matrix H_{ij}^{basic}, as described above, and are given a phylogenetic tree $\mathcal{B}$ with the targets at the leaves and a set $\mathcal{I}$ of internal nodes defining monophyletic target groups.

Since we want to detect unknown targets $t \notin \mathcal{T}$ to a degree that we can place them at an approximately correct location in the phylogenetic tree $\mathcal{B}$, we need to model a ‘‘typical’’ hybridization pattern of an unknown target that belongs to each particular monophyletic group.

Let v denote an internal node in $\mathcal{B}$ and let $L(v) := \{i : t_i \text{ is a leaf below } v\}$ denote the set of target indices that form the monophyletic group below v. Our approach is to postulate that probe p_i is ‘‘typical’’ for v if it hybridizes to more than half of the targets in $L(v)$. We thus define the hybridization vector $h(v)$ by the strict majority function,

$$h(v) \equiv (h_{v,1}, \ldots, h_{v,n}) \in \{0,1\}^n \text{ with}$$
$$h_{v,j} := 1 \Longleftrightarrow \sum_{i \in L(v)} H_{i,j}^{\text{basic}} > |L(v)|/2.$$

One alternative would be to use the logical and function (i.e., set $h_{v,j} := \wedge_{i \in L(v)} H_{ij}^{\text{basic}}$), but intuitively this does not capture the ‘‘typicality’’ of probes as well as the majority function. Nevertheless, other alternatives are certainly possible; the aim being to guess as precisely as possible the hybridization behavior of unknown targets in a monophyletic group, which is *per se* an impossible task.

To build the extended hybridization matrix H of size $(m + |\mathcal{I}|) \times n$, we define the first m rows as those in H^{basic}. To define the remaining $|\mathcal{I}|$ rows, we assign numbers $i(v)$ ranging from $m+1$ to $m + |\mathcal{I}|$ bijectively to the internal nodes $v \in \mathcal{I}$ and define the $i(v)$-th row of H as the majority vector $h(v)$.

An example of an extended hybridization matrix H with 5 targets and 10 probes, along with the phylogenetic target tree $\mathcal{B}$ with 4 internal nodes, is shown in Figure 1 (left).

2.4 Probabilistic hybridization model

As stated in Section 2.1, the expected hybridization result $r = r(T)$ of a target set $T \subset \{1, \ldots, m + |\mathcal{I}|\}$ is obtained by computing the logical `or` of the indicated rows of the hybridization matrix H. It is understood that if I contains representations of unknown targets u (indices ranging from $m+1$ to $m + |\mathcal{I}|$), r is not the actual hybridization pattern of T, since the actual behavior of u is unknown and only hypothesized to look similar to the corresponding row in H.

As an example, consider Figure 1 (right). The expected result for singleton target sets can be read directly from H (examples a, c). If $|T| \geq 2$, the result is the logical `or` of the corresponding rows (examples b, d–f). The set $\{v_{1,2}\}$ represents a *single typical unknown* target somewhere below $v_{1,2}$ (and no further targets) and must be distinguished from $\{1,2\}$ that consists of *two particular known* targets (and no further targets). Target sets may mix known and unknown targets (example d). Sometimes, the same result may occur for several distinct target sets (examples e, f; there are many more target sets giving rise to this ‘‘all ones’’ result). Other results may not be explainable by any target set at all without allowing errors (example g).

In order to model false positive and false negative hybridizations, we switch to a probabilistic model, where r becomes a random vector whose distribution depends on T and the assumed error rates. We use a model with two error parameters: f_- denotes the (per probe and target) probability that a hybridization fails, and f_+ denotes the (per probe) probability that a probe shows a signal although no hybridization should take place. In practice, we must assume error rates of up to 0.1.

We define $\mathcal{P}_i := \{j \in \{1, \ldots, n\} : H_{ij} = 1\}$ as the set of probes hybridizing to target i, and $\mathcal{T}_j := \{i \in \{1, \ldots, m + |\mathcal{I}|\} : H_{ij} = 1\}$ as the set of targets hybridizing to probe j.

For given T, in order to observe no signal at probe j, *all* of the $|T \cap \mathcal{T}_j|$ expected hybridizations must fail. Assuming independence between these failures, this event occurs with probability $f_-^{|T \cap \mathcal{T}_j|}$. Additionally, the probe must not show a false positive reaction; this event has probability $1 - f_+$ and is also assumed to be independent of potential failure events. It follows that

$$\eta_j(T) :\equiv \mathbb{P}(r_j = 0\,|\,T) = f_-^{|T\cap\mathcal{T}_j|}\cdot(1-f_+), \tag{1}$$

and that $\mathbb{P}(r_j = 1\,|\,T) = 1 - \eta_j(T)$.

We further assume that all probes react independently, such that the joint probability that the observed result is a particular vector $r = (r_j)$ is given by the product

$$\mathbb{P}(r\,|\,T) = \prod_{j=1}^{n} (1-\eta_j(T))^{r_j}\cdot(\eta_j(T))^{1-r_j}. \tag{2}$$

For example, assuming $f_+ = f_- = 0.05$, the result $r = (1,0,1,0,1,0,1,1,0,1)$ in Figure 1 (Example g) has probability $2.1\cdot 10^{-7}$ if $T = \{4\}$ and $1.3\cdot 10^{-8}$ if $T = \{\}$.

As an example on a larger scale, consider error rates of 10% in an experiment with 1000 probes and a target set T with a single target covered by 10 probes. We expect one false negative, nine true positive and 100 false positive probes. Even though the number of false positives is much larger than the number of true positives, correct target identification will be possible in most cases because the false positive probes do not paint a consistent picture, while the true positive probes do.

2.5 Target set prior

To identify a DNA sample, we are given a realization of r and are asked for the target set T that best explains the observation. In principle, we could proceed by a maximum likelihood approach, i.e., attempt to find T^* that maximizes $\mathbb{P}(r\,|\,T)$ over all T . However, from the example in Figure 1, we see that this would cause problems for results such as $r = (1,1,\ldots,1)$ that have many good explanations. In accordance with our sparseness assumptions and Occam's razor, we prefer a parsimonious explanation (small $|T|$), but the likelihood model specified by Eqs. (1), (2) actually prefers larger target sets.

We thus move to a Bayesian framework and introduce a prior probability distribution on the potential target sets, defined by a ''prevalence'' vector $f = (f_1,\ldots,f_{m+|\mathcal{I}|}) \in [0,1/2]^{m+|\mathcal{I}|}$, where f_i denotes the a-priori probability that target i is contained in T, and all target occurrences are assumed independent:

$$\mathbb{P}(T) = \prod_{i=1}^{m+|\mathcal{I}|} f_i^{T_i}\cdot(1-f_i)^{1-T_i}. \tag{3}$$

The relative magnitude f_i/f_k determines how much more likely it is a-priori to see target i in an unclassified sample than target k. Such ratios are available for many applications, e.g., the relative prevalences of HIV subtypes in patients. If nothing is known, a flat prevalence prior may be used where all f_i are equal. The absolute magnitude $F = \sum_i f_i$ should be chosen such that $f_i \ll 1/2$ for all i, and depending on how many probes are available to decide reliably on inclusion or exclusion of target i. In practice, we recommend $f_i \approx 0.01$ to favor non-inclusion of each target 99-fold over its inclusion a-priori.

2.6 Decoding hybridization results

Maximum a-posteriori. By Bayes Theorem, the posterior probability of a target set T given a hybridization result r is

$$\mathbb{P}(T\,|\,r) = \frac{\mathbb{P}(T)\cdot\mathbb{P}(r\,|\,T)}{\mathbb{P}(r)} \propto \mathbb{P}(T)\cdot\mathbb{P}(r\,|\,T), \tag{4}$$

where $\mathbb{P}(r)$ is a constant. We are interested in finding sets $T \subset \{1,\ldots,m+|\mathcal{I}|\}$ that explain r well in the sense that $\mathbb{P}(T\,|\,r)$ is high. For very small examples, such as the one in Figure 1, we can compute the posterior for all T directly and find the maximizing set T^* by brute force. For example, assuming error rates $f_+ = f_- = 0.05$ and prior prevalences $f_i = 0.33$ for all i, the two best explanations for the observation r in Figure 1 (Example g) are $T_1 = \{4\}$ with $\mathbb{P}(T_1\,|\,r) = 0.775$ and $T_2 = \{\}$ with $\mathbb{P}(T_2\,|\,r) = 0.094$.

However, since $\mathbb{P}(T\,|\,r)$ is a complicated function of T, direct maximization seems out of reach for realistically large datasets. Additionally, there may be several good distinct solutions.

Posterior marginals. For the above reasons, instead of maximizing the posterior, we estimate the *posterior marginals* $\mu_i := \mathbb{P}(T_i = 1\,|\,r)$ and the *posterior target set cardinality* $M := \mathbb{E}[|T|\,|\,r] = \sum_i \mu_i$ to decide how many and which targets are the best candidates for explaining r. In the toy example, we find that $\mu_4 = 0.81$ and $\mu_2 = 0.06$ are the highest posteriors and $M = 0.95$ indicates that we expect slightly less than one target to be present.

In larger problems, we estimate these quantities by Gibbs sampling from the posterior. The next paragraphs show that this can be done efficiently in our model.

Gibbs sampling. In our setting, Gibbs sampling consists of a pre-defined number of rounds, during each of which we update the target set T, which is initially random. Each round consists of $m + |\mathcal{I}|$ steps, and in step i of each round we decide whether target t_i should be included in or removed from T by considering the posterior ratio $\rho \equiv \rho_i(T)$ defined as follows: If $i \notin T$, let $T^+ := T \cup \{i\}$, otherwise, if $i \in T$, let $T^- := T\backslash\{i\}$, and let

$$\rho := \begin{cases} \mathbb{P}(T^+\,|\,r)/\mathbb{P}(T\,|\,r) & \text{if } i \notin T, \\ \mathbb{P}(T\,|\,r)/\mathbb{P}(T^-\,|\,r) & \text{if } i \in T. \end{cases}$$

In other words, ρ is the conditional posterior probability ratio of including and not including t_i in the target set, given the observation result r and the remaining components of the target set.

The update rule is then: If $i \notin T$, add i to T with probability $\mathbb{P}(T^+\,|\,r)/(\mathbb{P}(T^+\,|\,r) + \mathbb{P}(T\,|\,r)) = \rho/(\rho+1)$ (and leave T unchanged with the remaining probability $1/(\rho+1)$). If $i \in T$, remove it with probability $1/(\rho+1)$(and leave T unchanged with the remaining probability $\rho/(\rho+1)$).

In this way, we cycle through all targets in either a fixed or random order in each round. This defines an ergodic Markov chain on T with the posterior as stationary distribution, from which we sample the quantities of interest during S sampling rounds after W warmup rounds to allow for the Markov chain to converge towards its stationary distribution.

We estimate the posterior marginals as follows. In round τ when updating target i, remember the value $p_i^{(\tau)} := \rho/(\rho+1)$, where ρ is computed as described above. Then our estimate $\hat{\mu}_i$ for μ_i is $\hat{\mu}_i := \frac{1}{S}\sum_{\tau=W+1}^{W+S} p_i^{(\tau)}$, and our estimate for the target set size is $\hat{M} := \sum_{i=1}^{m} \hat{\mu}_i$.

Efficient computation of ρ-ratios. A key feature of this procedure is that the above ratios ρ can be efficiently computed in each step by taking advantage of the following observations.

Consider an update attempt $T \leftarrow T^+ = T \cup \{i\}$ with $i \notin T$, where, using Eqs. (1)–(3),

$$\begin{aligned}
\rho &= \frac{\mathbb{P}(T^+)}{\mathbb{P}(T)}\cdot\frac{\mathbb{P}(r\,|\,T^+)}{\mathbb{P}(r\,|\,T)} \\
&= \frac{f_i}{1-f_i}\cdot\prod_{j\in\mathcal{P}_i}\left(\frac{1-\eta_j(T^+)}{1-\eta_j(T)}\right)^{r_j}\cdot\left(\frac{\eta_j(T^+)}{\eta_j(T)}\right)^{1-r_j} \\
&= \frac{f_i}{1-f_i}\cdot\prod_{j\in\mathcal{P}_i}\begin{cases} f- & \text{if } r_j = 0, \\ \frac{1-\eta_j(T)\cdot f_-}{1-\eta_j(T)} & \text{if } r_j = 1 \end{cases} \\
&= \xi_i\cdot\prod_{\substack{j\in\mathcal{P}_i \\ r_j=1}}\frac{1-\eta_j(T)\cdot f_-}{1-\eta_j(T)},
\end{aligned}$$

where $\xi_i := \frac{f_i}{1-f_i}\cdot f_-^{|\{j\in\mathcal{P}_i : r_j=0\}|}$. Note that in the prior ratio, everything except the i-th term cancels out, and in the likelihood ratio, all terms related to probes that do not hybridize to the i-th target also cancel out. The prior ratio and probability of necessarily false negative probes to include t_i in the target set is summarized in the factor ξ_i. Similarly, for an update attempt $T \leftarrow T\backslash\{i\}$, we have

$$\rho = \xi_i\cdot\prod_{\substack{j\in\mathcal{P}_i \\ r_j=1}}\frac{1-\eta_j(T)}{1-\eta_j(T)/f_-}.$$

The ξ_i can be pre-computed and never change during the sampling phase, and the remaining product generally has few terms: the relevant probe

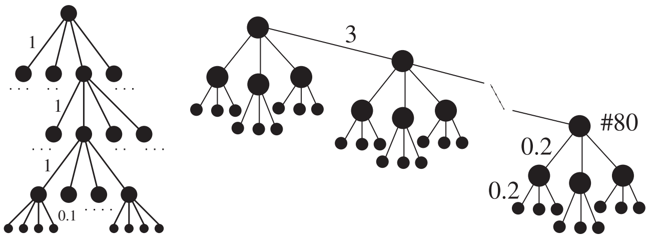


Fig. 2. *Left*: Balanced treemodel with 256 leaves. *Right*: Cherry tree model with 720 leaves.

Table 1. Summary statistics of the datasets. #Known refers to the number of known targets in the dataset, #Probes is the number of probe candidates selected by PROMIDE. #Hybs is the number of 1s in the hybridization matrix H^{basic}. The average number of hybridizations per probe and per target is shown in the next two columns. Finally, #Unknown denotes the number of unknown targets. Numbers are averages over the dataset instances

Name	Known	Probes	Hybs	H/probe	H/target	Unknown
`bal`	181	4038	10557	2.61	58.2	75
`cher`	539	8536	24485	2.87	45.4	181
`meio`	302	8837	16439	1.89	54.5	56

set $\mathcal{P}_i \cap \{j : r_j = 1\}$ can also be precomputed for every target i and will generally be sparse.

To evaluate the ratios within the products quickly, we maintain and update the vector $\eta = (\eta_j(T))_{j=1,\dots,n} = \mathbb{P}(r_j = 0 \mid T)$ as defined in Eq. (1); in fact, we only require the elements η_j for which $r_j = 1$. Initially, T is empty, and $\eta_j(T) = 1 - f_+$ for all probes j. When T is enlarged to $T \cup \{i\}$ (resp. reduced to $T \setminus \{i\}$), we update $\eta_j \leftarrow \eta_j \cdot f_-$ (resp. $\eta_j \leftarrow \eta_j / f_-$) for all $j \in \mathcal{P}_i$ with $r_j = 1$.

3 EVALUATION

3.1 Datasets

We evaluate the proposed method on one biological dataset of organisms from the Meiobenthos and on two simulated datasets. Summary statistics of the datasets are shown in Table 1. The simulated datasets were generated with the REFORM (Random Evolutionary FORest Model) software [Rahmann, 2003b], freely available at http://gi.cebitec.uni-bielefeld.de/people/rahmann, that applies an evolutionary Markov process along a phylogenetic tree (specified in a small modeling language) to a random root sequence.

Simulated dataset `bal`. We generate 256 targets (leaf sequences) from a balanced tree as shown in Figure 1 (left). The tree has four levels below the root, and each internal node has out-degree four. For the internal branches, the evolutionary time is 1 percent of expected mutations (PEM), for the branches to the leaves, it is 0.1 PEM. Additionally, there are small insertion and deletion probabilities (details not shown). This leads to target sequence lengths between 970 and 1030, generated from a root sequence of length 1000. In order to have both known and unknown targets available, we traverse the tree top-down and prune the second and third child of each (internal or leaf) node we encounter with 20% probability. We generate 8 instances of this dataset with different random root sequences and random prunings. This leads to 146–210 known targets.

Simulated dataset `cher`. The tree consists of 80 nodes arranged in a linear chain with an inter-node distance of 3 PEM; see Figure 1 (right). Each chained node has three children in addition to the next node in the chain at distance 0.2 PEM, and each of these has in turn three children at the same distance. From the visual impression of this tree topology, we call this the cherry tree model. The 720 targets are generated from a root sequence of length 600, and their length ranges between 580 and 620. To generate unknown targets, the second child of each node is pruned away from the tree with 40% probability, leading to 527–555 known targets in the 8 generated instances of the dataset.

Real dataset `meio`. We use a set of 358 28S rDNA sequences from different organisms present in the Meiobenthos related by a phylogenetic tree [Markmann, 2000]. The set contains redundancies and many close homologs and finding unique probes is difficult [Schliep *et al.*, 2003, Kaderali and Schliep, 2002]. To generate unknown targets, we remove the the last leaf child of an internal node (if more than one exists) with 50% probability. We generated 5 instances of this dataset; in each distance, a different random target set is removed from the tree (cf. Table 1).

Probe selection. After randomly separating the sequences into known and unknown targets as described above, we use PROMIDE to select short oligonucleotide probes for the known targets. We pick all group-specific (groups were restricted to be of size 50 or below) 19–21-mers with Gibbs free energy between -20 and -19.5 kcal/mol at 40°C and a salt correction parameter of -2.6, according to the model parameters from [SantaLucia, 1998]. We create the extended hybridization matrix of all known targets against all probes, as described in Section 2.3.

We emphasize that the unknown targets have no influence on the probe selection process, but after the probes have been determined, we can of course compute their hybridization patterns. Although here we might face the problem of unclear signals (long common substrings), we take the approach that only exact full-length probe-target matches lead to a signal. The possibility of weaker cross-hybridization signals is handled by a correspondingly high false-positive error rate in our error model (up to $f_+ = 0.10$), see below.

3.2 Hybridization simulations and decoding

Simulations. We performed simulations of hybridization experiments to estimate the efficiency of our approach in detecting both known and unknown targets. We randomly sample target sets which are taken as the true result of the experiment. The sampling strategy is different for sets of known targets, for unknown targets, and mixed sets.

(1) `known`: We attempt to correctly detect the empty target set $T = \{\}$ and each of the $|\mathcal{T}|$ singleton sets $T = \{t_i\}$, $i = 1, \dots, m$, where m varies for each dataset instance. For target sets cardinalities $2, \dots, 6$, we sample 500 random sets each.

(2) `unknown`: For each unknown target (each removed leaf from the original phylogenetic tree), we determine its lowest existing ancestor in the remaining tree; this is an internal node. As discussed above, we take this node as a representative of any

Table 2. Average fraction of correctly identified true $|T|$ targets (hits) among the $|T|$ top randed targets given by the decoder for different datasets (rows) and different types of datasets (columns). For unknown and mixed datasets, a target is counted as a hit if either either the internal node representing the unknown target (colums ''Exact''), or taking a broader view, the node or its direct children (columns titled ''Fam.'' for family) are detected

		known	unknown		mixed	
Name	$f_+ = f_-$		Extract	Fam.	Exact	Fam.
bal	0.05	0.98	0.38	0.69	0.80	0.89
	0.1	0.94	0.36	0.68	0.77	0.86
cher	0.05	0.97	0.11	0.51	0.77	0.84
	0.1	0.94	0.11	0.54	0.71	0.83
meio	0.05	0.97	0.08	0.45	0.71	0.83
	0.1	0.96	0.06	0.44	0.70	0.82

unknown target in the subtree below it. Therefore, ideally, this node is the target that we would like to detect, although the hybridization pattern of the unknown target will generally differ from the ''majority vote'' pattern of the internal node. Also, different unknown targets may map to the same node. Because of these inherent difficulties, we only attempt to detect a single unknown target.

(3) mixed: Finally, we attempt mixed sets with exactly one unknown target and between 1 and 3 known targets. For each cardinality, 500 random sets are sampled.

For each target set $T \subset \{1, \ldots, m + |\mathcal{I}|\}$, we simulate 10 independent hybridization results according to the error model described in Section 2.4, i.e., for each probe p_j, we determine the number of targets in T to which p_j would hybridize and let each hybridization fail independently with probability f_-; finally, there is a probability of f_+ that p_j shows an unspecific positive signal. This simulation was performed once with error rates $f_+ = f_- = 0.05$ and again with $f_+ = f_- = 0.1$.

Decoding. We ran our own TPDC decoding software with a uniform prior $f = (f_i)$, $i = 1, \ldots, m + |\mathcal{I}|$ on all targets such that $\sum_i f_i = 3$. The error parameters $f_- = f_+ \in \{0.05, 0.1\}$ were the same as used in the simulations. In practice, the error rates are not known and must be estimated. After 200 warmup rounds, the marginal target posteriors were estimated from the subsequent 2000 rounds; these values were found to be sufficiently accurate when compared to substantially longer runs. The output consists of a list of targets sorted by marginal posterior and additional diagnostics. Only targets with a posterior exceeding 0.001 were included in further analysis.

3.3 Results

Ideally, we observe exactly the true targets as the top entries of the list returned by the decoder. Depending on the similarity of hybridization patterns and on the noise level, we must expect a number of high-posterior targets that do not belong to the target set. In some of those cases, the ''offender'' is likely to be a close relative of the true target. We take this fact into account in our evaluation.

The success rates of our approach for a total of about 2,292,380 simulated experiments are summarized in Table 2. Simulation results for the datasets bal, cher are averaged over the 8 instances, for meio over the 5 instances, over all target set cardinalities, and over the 10 repetitions of simulated hybridizations.

Our method is able to correctly identify over 94% of known targets in simulated experiments with realistic error rates. If there are neither known nor unknown targets present, the maximal target posterior observed in all repetitions and data sets was 0.15 and posteriors exceeding the 0.001 posterior threshold were predominantly (over 95%) below 0.01, implying a negligible false positive rate. The results for unknown targets suggest that our simple approach for defining the hybridization pattern of its parent is not sufficient. There is a jump in performance when also direct children are counted as a hit. Then, up to 70% of the unknowns were correctly assigned to their clade in the complete tree. Detailed summaries for the 2,292,380 simulated experiments are found in the supplementary material.

4 DISCUSSION

We present an approach for decoding hybridizations experiments when targets are related by a phylogenetic tree and non-unique oligonucleotide probes are used in a statistical group testing setting. Hybridization patterns of internal nodes of the tree are obtained from leaves based on a majority rule as typical patterns for unknown targets in the respective subtree. A Bayesian framework combined with a Markov chain Monte Carlo approach allows efficient and robust estimation of target posterior marginals.

Our method correctly identifies over 94% of known targets, and about 45% to 70% of unknown targets were correctly assigned to their clades in the phylogenetic tree. The lower figures for unknown targets are explained by the fact that the majority-vote hybridization patterns of the internal nodes do not (and cannot) match exactly the hybridization patterns of unknown targets.

We found that our estimate of the target set size $|T|$ matches the true value in virtually all of the cases when rounded to the nearest integer. It follows that the rate of falsely identified targets is between 2% and 6% for known targets.

More detailed analysis of the high-ranking targets may improve the resolution of the method in the presence of unknowns, as we correctly identify clades but do not provide a statistical test for the hypothesis that unknowns belonging to this clade are present.

In a practical application of the method, the true target set size $|T|$ and the error rates f_+, f_- for the decoding procedure will be unknown. However, we can estimate $|T|$ by the sum of the posterior marginals, and our results show that the method is robust, even for relatively high error rates, which makes it reasonable to use with slight overestimates of error-rates, possibly at the expense of less pronounced posterior magnitudes. For the robustness of the method, a high probe coverage per target is necessary, and future work may show to which degree the probe set may be reduced without affecting our ability to detect unknown targets too severely.

Our results on biological and simulated data demonstrate that we can cope effectively with the incomplete phylogenies available in practical applications and that the method is robust with respect to evolution of targets between time of design and time of experiment. We are not aware of previous studies that consider the problem of recognizing unknown or fast evolving targets in such a manner.

ACKNOWLEDGEMENTS

Many thanks to Diethard Tautz and Win Hide for helpful discussions. We also thank Jonas Heise for help with preprocessing the Meiobenthos dataset.

REFERENCES

L. Kaderali and A. Schliep. Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, 18(10):1340–1349, Oct 2002.

G. W. Klau, S. Rahmann, A. Schliep, M. Vingron, and K. Reinert. Optimal robust non-unique probe selection using Integer Linear Programming. *Bioinformatics*, 20(Suppl 1):i186–i193, Aug 2004.

M. Markmann. *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, 2000.

C. Putonti, S. Chumakov, R. Mitra, G. E. Fox, R. C. Willson, and Y. Fofanov. Human-blind probes and primers for dengue virus identification. *FEBS J*, 273(2):398–408, Jan 2006.

S. Rahmann. Rapid large-scale oligonucleotide selection for microarrays. In *Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB)*, pages 54–63. IEEE, 2002.

S. Rahmann. Fast large scale oligonucleotide selection using the longest common factor approach. *Journal of Bioinformatics and Computational Biology*, 1(2):343–361, 2003a.

S. Rahmann. REFORM (Random Evolutionary FORests Modeling software), 2003b.

S. Rash and D. Gusfield. String barcoding: Uncovering optimal virus signatures. In *Proceedings of RECOMB 2002*, pages 254–261, April 2002.

J. SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the U.S.A.*, 95:1460–1465, 1998.

A. Schliep, D. C. Torney, and S. Rahmann. Group testing with DNA chips: Generating designs and decoding experiments. In *Proceedings of the 2nd IEEE Computer Society Bioinformatics Conference (CSB 2003)*, pages 84–93. IEEE, 2003.

D. Wang, A. Urisman, Y.-T. Liu, M. Springer, T. G. Ksiazek, D. D. Erdman, E. R. Mardis, M. Hickenbotham, V. Magrini, J. Eldred, J. P. Latreille, R. K. Wilson, D. Ganem, and J. L. DeRisi. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol*, 1(2):E2, Nov 2003.