Reconstructing distances in physical maps of chromosomes with nonoverlapping probes

John Kececioglu* Sanjay Shete† Jonathan Arnold‡

Abstract

We present a new method for reconstructing the distances between probes in physical maps of chromosomes constructed by hybridizing pairs of clones under the so-called sampling-without-replacement protocol. In this protocol, which is simple, inexpensive, and has been used to successfully map several organisms, equal-length clones are hybridized against a clone-subset called the *probes*. The probes are chosen by a sequential process that is designed to generate a pairwise-nonoverlapping subset of the clones. We derive a likelihood function on probe spacings and orders for this protocol under a natural model of hybridization error, and describe how to reconstruct the most likely spacing for a given order under this objective using continuous optimization. The approach is tested on simulated data and real data from chromosome VI of *Aspergillus nidulans*. On simulated data we recover the true order and close to the true spacing; on the real data, for which the true order and spacing is unknown, we recover a probe order differing significantly from the published one. To our knowledge this is the first practical approach for computing a globally-optimal maximum-likelihood reconstruction of interprobe distances from clone-probe hybridization data.

Keywords Computational biology, physical mapping of chromosomes, sampling without replacement protocol, maximum likelihood, convex optimization

1 Introduction

Physical mapping in molecular biology is the task of reconstructing the order and location of features of biological interest along a chromosome. The features may be sites at which restriction enzymes cut, so-called sequence-tagged sites that are identified by short, uniquely-occurring sequences, or positions of clones that contain fragments of the chromosome. There is a diverse array of approaches for constructing maps of such features depending on the type of data that is collected, including mapping by nonunique probes [2, 18], mapping by unique probes [1, 11, 12], mapping by unique endprobes [7], mapping by nonoverlapping probes [8], mapping from restriction-fragment length data [10, 13], radiation-hybrid mapping [24, 5], and optical mapping [21, 14, 16]; there are many probabilistic analyses of various approaches [15, 4, 28, 27, 26]; and a wide variety of computational techniques have been employed or suggested, including greedy algorithms [18], simulated annealing [20, 25, 2, 1], linear programming [7, 12, 8], and semidefinite programming [6].

In this paper we develop a maximum-likelihood approach for a type of physical mapping known as the sampling-without-replacement protocol. The protocol is inexpensive, simple to carry out in the lab, and uses widely-available technology. Organisms that have been mapped with this technique include *Schizosaccharomyces pombe* [19], *Aspergillus nidulans* [22], and *Pneumocystis carinii* [3]; mapping projects in progress using the technique include *Neurospora crassa* and *Aspergillus flavus*.

In the protocol, a library $\mathcal{L}$ consisting of many overlapping clones that each sample a fragment of the chromosome is developed. Clones in $\mathcal{L}$ are size-selected to have a target length, and are arrayed on a plate. A subset of the clones called the *probe set* $\mathcal{P}$ is then obtained by the following sequential process. Initially, $\mathcal{P} = \emptyset$ and $\mathcal{S} = \mathcal{L}$. At the ith iteration of the process, choose a clone P_i from $\mathcal{S}$ at random, remove P_i from $\mathcal{S}$, and add it to $\mathcal{P}$. Hybridize P_i against all the clones in the library by extracting complementary DNA from both of its ends and washing the DNA over the arrayed plate, recording all clones in the library to which the DNA sticks. Remove from $\mathcal{S}$ all clones in the library that have a positive hybridization result with P_i. Then repeat this process for the next iteration, stopping once $\mathcal{S}$ becomes empty.

We call the final set $\mathcal{P}$ the *probe set*, and the set $\mathcal{C} = \mathcal{L} - \mathcal{P}$ the *clone set*. The results of the experiments are summarized in a probe-clone hybridization matrix H that records the outcomes of all hybridizations between the probes in $\mathcal{P}$ and the clones in $\mathcal{C}$.

Notice that if a clone $C_i \in \mathcal{C}$ overlaps with a probe $P_j \in \mathcal{P}$ in the chromosome, it must overlap with one of the ends of P_j, as all probes and clones are of the same length. Such an overlap corresponds to a portion of DNA that is in com-

*Corresponding author. Department of Computer Science, University of Georgia, Athens, GA 30602-7404. Email: kece@cs.uga.edu.
Research supported by National Science Foundation CAREER Award DBI-9722339.
†Department of Statistics, University of Georgia, Athens, GA 30602. Email: sanjay@stat.uga.edu
‡Department of Genetics, University of Georgia, Athens, GA 30602. Email: arnold@genetics.uga.edu

mon between the clone and the end of the probe. In the absence of error, the complementary DNA from the end of P_j will stick to C_i, and the hybridization test of P_j versus C_i will be a positive result; thus clone C_i will be removed from set $\mathcal{S}$ at the jth iteration. This implies that in the absence of error the probe set $\mathcal{P}$ is a maximal nonoverlapping subset of the library.

Suppose that in hybridization matrix H enough of the clone-probe overlap structure is represented that we can recover the order of the probes $\mathcal{P}$ across the chromosome. Then for every consecutive pair of probes P and Q in this order, we can examine H for the presence of a *linking clone* C that overlaps with both P and Q. The probe set $\mathcal{P}$ together with a linking clone for every consecutive pair forms a minimal set of clones that cover the chromosome. A map giving the order of the probes across the chromosome is then very useful, since by individually sequencing just the probes and linking clones and overlapping the sequences in the order given by the map, we can reconstruct the DNA sequence of the chromosome.

In reality, hybridization tests do not perfectly record the overlap structure of probes and clones. Hybridization results contain random false positives and false negatives. A probe can also hybridize to a nonoverlapping clone due to repeated DNA in the chromosome. In general, clones can be chimeric, which means they sample two or more fragments of the chromosome, and can contain deletions, which happens when portions of the DNA get spliced out during cloning. In the mapping projects using this protocol at the University of Georgia, however, clones are produced by cosmids, which are small enough that chimerism and deletions are not a significant problem. In our treatment we model false positives and false negatives, but not chimerism, deletions, or repeats. Hence false hybridizations due to repeats are treated as a series of isolated false positives.

Related work Prior work on mapping by the sampling-without-replacement protocol, by Cuticchia, Arnold and Timberlake [9], Wang, Prade, Griffith, Timberlake and Arnold [25], and Mott, Grigoriev, Maier, Hoheisel and Lehrach [20], has largely used local-search heuristics such as simulated annealing to try to find a probe order that minimizes the Hamming-distance traveling-salesman objective. While minimizing this objective is not known to optimize any natural measure of the goodness of a map, Xiong, Chen, Prade, Wang, Griffith, Timberlake and Arnold [27] have shown that under certain assumptions on the distribution of clones, the Hamming-distance objective is statistically consistent; this means that as the number of clones goes to infinity, an exact algorithm for the Hamming-distance traveling salesman problem would recover the correct probe order with probability one.

Christof and Kececioglu [8] recently showed that the problem of computing a maximum-likehood probe *order* in the sampling-without-replacement protocol in the presence of false-positive and -negative hybridization error can be reduced to the problem of finding the minimum number of ones to change to zeroes in hybridization matrix H so that the resulting matrix H' has at most 2 ones per row and the consecutive-ones property on rows. They then showed how to formulate this problem as an integer linear program, and developed a branch-and-cut algorithm for computing an optimal maximum-likelihood probe order. Using this approach, they were able to compute optimal probe orders for realistic-sized instances on simulated data, and probe orders with significantly fewer false positives on real data than the best-possible map obtainable by a Hamming-distance traveling-salesman approach. In this paper we complement the work in [8] by developing a practical method for computing a globally-optimal maximum-likelihood reconstruction of the interprobe *distances*, given a probe order.

Plan of the paper In the next section we give a maximum likelihood formulation of the problem of mapping by the sampling-without-replacement protocol in the presence of false positive and false negative error, which we call *Mapping by Nonoverlapping Probes*. The problem is unique in that the goal is to reconstruct the most likely order and *spacing* of probes along the map from the hybridization data. Section 3 then derives the likelihood function on probe orders and spacings for this formulation, which has a remarkably simple closed form. Section 4 explains how we tackle the maximization of this function for a fixed probe order using continuous optimization. Section 5 presents results of some experiments with a preliminary implementation of this approach. We then conclude with several directions for further research.

2 The problem

In our maximum likelihood formulation we do not model the sequential process of choosing the probes, and hence we operate under the assumption that the probes form a nonoverlapping set. We write $\{P_1, \ldots, P_n\}$ for the set of n probes, $\{C_1, \ldots, C_m\}$ for the set of m clones, and we formulate the problem as follows.

The task is to recover the probe order $\pi = (\pi_1 \; \pi_2 \; \cdots \; \pi_n)$ and the probe spacing $x = (x_1 \; x_2 \; \cdots \; x_n)$ as illustrated in Figure 1, given the $m \times n$ clone-probe hybridization matrix H containing false positive and false negative errors. Permutation π gives the names of the probes in left-to-right order across the chromosome. Vector x gives the distance beween consecutive probes, where component x_j is the distance between the left end of P_{π_j} and the right end of $P_{\pi_{j-1}}$. Matrix $H = (h_{ij})$ is a 0-1 matrix, with

$$h_{ij} \;=\; \begin{cases} 1, & P_j \text{ hybridizes to } C_i; \\ 0, & \text{otherwise.} \end{cases}$$

We assume that all clones are the same length, that the probes are nonoverlapping, and that we know

- L, the length of the chromosome,
- ℓ, the length of a clone,
- ρ, the probablility that an entry of H has been corrupted into a false positive, and
- η, the probability that an entry of H is a false negative.

As stated, this is not a well-posed problem. In the presence of false positives and negatives, any permutation π of $\{1, \ldots, n\}$ and any positive vector x for which $\sum_{1 \le i \le n} x_i \le L - n\,\ell$ are an explanation of the data. To obtain a well-defined problem, we invoke the principle of maximum likelihood, which says that the best reconstructed map is that π and x that are most likely to have given rise to H. If we write $p(\pi, x \mid H)$ for the probability that π and x are the true order and spacing given the observed matrix H, a *maximum likelihood reconstruction* is a pair (π^*, x^*) that maximizes $p(\pi, x \mid H)$. We take the following as our definition of the problem.

Definition 1 (Mapping by Nonoverlapping Probes) The *Mapping by Nonoverlapping Probes Problem* is the following. The input is the clone-probe hybridization matrix H, the chromosome length L, the clone length ℓ, the false positive probability ρ, and the false negative probability η. The output is a probe order and probe spacing pair (π, x) that maximize $p(\pi, x \mid H)$ under the assumption that the probes are a collection of nonoverlapping clones, all clones are of equal length, that the left ends of clones are uniformly distributed across the chromosome, and that the entries of H have been independently corrupted with false positive probability ρ and false negative probability η. □

We can derive the function $p(\pi, x \mid H)$ using Bayes' theorem:

$$p(\pi, x \mid H) \quad = \quad \frac{p(H \mid \pi, x)\; p(\pi, x)}{p(H)}.$$

In this equation, $p(H \mid \pi, x)$ is the probability of observing H given that π and x are the true order and spacing, $p(\pi, x)$ is the probability that π and x occur in nature, and $p(H)$ is the probability of observing H.[1] Since $p(H) = \sum_{\widetilde{\pi}} \int_{\widetilde{x}} p(H \mid \widetilde{\pi}, \widetilde{x})\, p(\widetilde{\pi}, \widetilde{x}) d\widetilde{x}$, the denominator is a constant independent of π and x and can be ignored. Since the names given to probes and the spaces between probes are independent, $p(\pi, x) = p(\pi)p(x)$. Since names are assigned to probes completely randomly, $p(\pi) = 1/n!$, which is independent of π and can also be ignored. Thus the only relevant quantities are $p(H \mid \pi, x)$ and $p(x)$.

If the probability density function $p(x)$ on probe spacings is uniform, this factor can be ignored as well. For the model considered below, we do not yet know the density function $p(x)$, but it does not appear to be uniform. We concentrate instead on deriving the function $p(H \mid \pi, x)$, and take maximizing it as our objective. This will differ from truly maximizing $p(\pi, x \mid H)$ according to the bias due to $p(x)$.

We next derive function $p(H \mid \pi, x)$ under the simplest process by which H can be generated from π and x with false positives and negatives. This process has three stages:

(1) each clone is thrown down uniformly and independently across the chromosome,

(2) for the row of the hybridization matrix corresponding to a given clone, the probes that a clone overlaps get a one in their column, and zeros are placed everywhere else, and

(3) the ones and zeros are corrupted randomly and independently with probability η and ρ respectively.

3 The objective function

To derive $p(H \mid \pi, x)$ under this model, notice that each row of H is independent of the other rows, since each clone is thrown down independently and each entry is independently corrupted. Writing H_i for the ith row of H then, it suffices to work out $p(H_i \mid \pi, x)$, since

$$p(H \mid \pi, x) \quad = \quad \prod_{1 \le i \le m} p(H_i \mid \pi, x).$$

To derive $p(H_i \mid \pi, x)$, notice that in the absence of error there are only three possible types of overlaps that can occur with a given clone C_i as illustrated in Figure 2:

(1) Clone C_i overlaps with no probe. If the left end of clone C_i falls between the left ends of probes $P_{\pi_{j-1}}$ and P_{π_j} but C_i overlaps with *neither* P_{π_j} nor $P_{\pi_{j-1}}$, we write $C_i \in N_j^\pi$. (If C_i falls to the left of P_{π_1} but does not overlap with it, we write $C_i \in N_1^\pi$, and if C_i falls to the right of P_{π_n} but does not overlap with it, we write $C_i \in N_{n+1}^\pi$.)

(2) Clone C_i overlaps with exactly one probe. If it overlaps with *only* probe P_{π_j}, we write $C_i \in O_j^\pi$.

(3) Clone C_i overlaps with exactly two probes. If it overlaps with *both* probe P_{π_j} and $P_{\pi_{j+1}}$, we write $C_i \in B_j^\pi$.

In Appendix A, we derive $p(H_i \mid \pi, x)$ by summing over the disjoint events $C_i \in N_j^\pi$, $C_i \in O_j^\pi$, and $C_i \in B_j^\pi$. For here, note that the domain $\mathcal{S}$ of the probe order permutation π is the set of all permutations on $\{1, \cdots, n\}$, and the domain $\mathcal{D} \subseteq \mathcal{R}^n$ of the spacing vector x is the set

$$\mathcal{D} \quad := \quad \Big\{ (x_1, \cdots, x_n) \in \mathcal{R}^n \;:\; \text{each } x_i \ge 0, \text{ and } \quad L - n\ell - \sum_{1 \le i \le n} x_i \;\ge\; 0 \Big\}. \qquad (1)$$

We summarize the derivation in the following theorem.

Theorem 1 (Objective function) *For hybridization matrix H, let $f : \mathcal{S} \times \mathcal{R}^n \to \mathcal{R}$ be*

$$f(\pi, x) \quad := \quad - \sum_{1 \le i \le m} \ln \Big(a_i^\pi \; - \sum_{1 \le j \le n+1} b_{ij}^\pi \; \min\{x_j, \ell\} \Big),$$

where the coefficients a_i^π and b_{ij}^π are given by Equations (2) through (5) in the Appendix, and we define

$$x_{n+1} \quad := \quad L - n\ell - \sum_{1 \le i \le n} x_i.$$

Then for a fixed probe order π,

$$\operatorname*{argmax}_{x \in \mathcal{D}} \; p(H \mid \pi, x) \quad = \quad \operatorname*{argmin}_{x \in \mathcal{D}} \; f(\pi, x),$$

where $\mathcal{D}$ is given by Equation (1). □

In other words, if we can evaluate the following objective function on permutations,

$$g(\pi) \quad := \quad \min_{x \in \mathcal{D}} f(\pi, x),$$

(and recover the minimizing x for a given π), we can reduce the continuous problem of maximizing $p(H \mid \pi, x)$ to a discrete search for a permutation that minimizes $g(\pi)$.[2] We now describe how we tackle the evaluation of $g(\pi)$.

[1] Since π and x are values taken on by underlying random variables Π and X, when we write $p(\pi, x)$ this is shorthand for $p(\Pi = \pi, X = x)$. Furthermore, since π is a discrete variable while x is a continuous variable, when we write $p(\pi, x)$ this is the joint probability density function of a discrete and a continuous random variable evaluated at π and x.

[2] Note that this does not solve the problem of finding a pair (π, x) that maximizes $p(\pi, x \mid H)$: the objective $f(\pi, x)$ is missing a term of $-\ln p(x)$, as we do not know the density function $p(x)$.

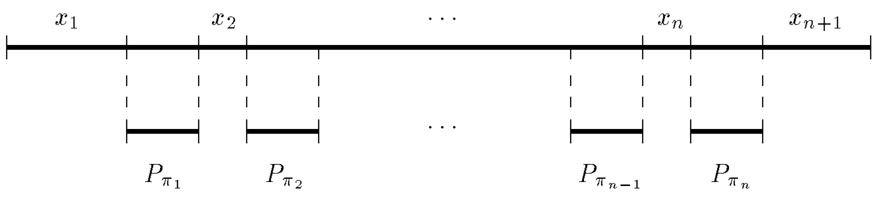

Figure 1 The problem is to reconstruct the probe order permutation $\pi = (\pi_1\ \pi_2\ \cdots\ \pi_n)$ and the probe spacing vector $x = (x_1\ x_2\ \cdots\ x_n)$ from the clone-probe hybridization matrix H. The probe set $\{P_1, P_2, \cdots, P_n\}$ is chosen to form a non-overlapping subset of the clones. Clones are size-selected to have the same length.

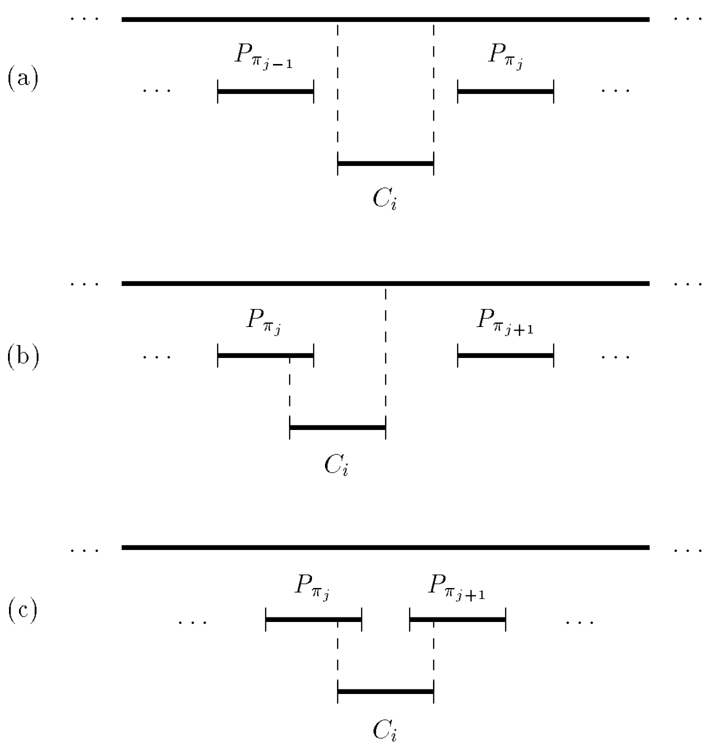

Figure 2 The three possible types of clone-probe overlaps. (a) $C_i \in N_j^\pi$. (b) $C_i \in O_j^\pi$. (c) $C_i \in B_j^\pi$.

4 Evaluating the objective for a fixed permutation

In this section, for a fixed π let us we write $f(x)$ for $f(\pi, x)$, and define

$$f_i(x) \quad := \quad a_i^\pi \; - \sum_{1 \le j \le n+1} b_{ij}^\pi \; \min\{x_j, \ell\}.$$

Then

$$f(x) \quad = \quad - \sum_{1 \le i \le m} \ln f_i(x).$$

Below we show that f is convex in certain convex regions of $\mathcal{D}$, so that a greedy procedure such as gradient descent will find the global minimum of f in such a region. We describe how we choose these regions of $\mathcal{D}$, and then explain how to find the direction of greatest decrease in f in such a constrained region for the gradient descent procedure. A very readable summary of the facts from optimization that we use is given by Lengauer [17].

4.1 Convexity

Recall that a set $\mathcal{C} \subseteq \mathcal{R}^n$ is a *convex set* if for all points p and q in $\mathcal{C}$ and all $0 \le \lambda \le 1$, the point $\lambda p + (1-\lambda) q$ is in $\mathcal{C}$. A function $h : \mathcal{C} \to \mathcal{R}$ defined on a convex set $\mathcal{C}$ is a *convex function* if for all points p and q in $\mathcal{C}$ and all $0 \le \lambda \le 1$,

$$h\big(\lambda p \; + \; (1-\lambda) q\big) \quad \le \quad \lambda \, h(p) \; + \; (1-\lambda) \, h(q).$$

Informally, a convex function is bowl-shaped.

Let us call a region $\mathcal{C} \subseteq \mathcal{D}$ *good* if for all points $x \in \mathcal{C}$ and all $1 \le j \le n+1$, $x_j \ne \ell$, where x_{n+1} is defined as in Theorem 1. The relevance of good regions is that they are the regions throughout which $f(x)$ is differentiable.

In a good region $\mathcal{C}$ consider all points $x = \mathbf{p} + s\mathbf{v}$ for $s \ge 0$, which is the ray traced by moving from point $\mathbf{p} \in \mathcal{C}$ in direction $\mathbf{v} = (v_1, \ldots, v_n)$. Along such a ray the derivative of f is well-defined and is equal to

$$\frac{d}{ds} f(x) \quad = \quad - \sum_{1 \le i \le m} \frac{1}{f_i(x)} \, \frac{d}{ds} f_i(x),$$

where

$$\frac{d}{ds} f_i(x) \quad = \quad \sum_{1 \le j \le n} v_j \Big(b_{ij}^\pi \; u(x_j) \; - \; b_{in+1}^\pi \; u(x_{n+1}) \Big),$$

and where $u(\cdot)$ denotes a unit step function at ℓ:

$$u(x) \quad := \quad \begin{cases} 1, & x < \ell; \\ \bot, & x = \ell; \\ 0, & x > \ell. \end{cases}$$

Taking a second derivative along the ray yields

$$\frac{d^2}{ds^2} f_i(x) \quad = \quad 0,$$

so that

$$\frac{d^2}{ds^2} f(x) \quad = \quad \sum_{1 \le i \le m} \left(\frac{1}{f_i(x)} \, \frac{d}{ds} f_i(x) \right)^2 \quad \ge \quad 0.$$

This implies that in every convex region $\mathcal{C} \subseteq \mathcal{D}$ that is good, function f is convex.

A key property of convex functions is that a local minimum of a convex function f in a convex set $\mathcal{C}$ is a *global minimum* of f on $\mathcal{C}$ [17]. Thus if we can divide $\mathcal{D}$ into a small number of good convex regions, it suffices to apply in each region an algorithm that is only guaranteed to find a local minimum; the best of these local minima is the global minimum of f over the regions.

Define

$$\begin{aligned} \mathcal{D}_{ab} \quad := \quad \big\{ x \in \mathcal{D} \; : \; & a\, x_1 \le a\, \ell, \\ & x_i \le \ell \text{ for all } 1 < i < n{+}1, \text{ and} \\ & b\, x_{n+1} \le b\, \ell \big\}, \end{aligned}$$

and consider the four regions $\mathcal{D}_{+1+1}$, $\mathcal{D}_{+1-1}$, $\mathcal{D}_{-1+1}$, and $\mathcal{D}_{-1-1}$. These regions correspond to constraining all interior distances between probes to be at most ℓ, and then forcing the exterior distances x_1 and x_{n+1} to be on one side of ℓ. Each region is an intersection of halfspaces, and hence is a convex set. The interior of each is a good region, and for any ray originating in the interior we can make the appropriate choice for the derivative at the boundary so that the derivative along the ray is continuous throughout the region. Thus we can find the global minimum in each of these four regions by gradient descent as described below.

This does not necessarily find the global minimum of f on $\mathcal{D}$. However, notice that for our function f, if a spacing vector x is modified by trading distance between two components $x_i \ge \ell$ and $x_j \ge \ell$ in such a way that both remain at least ℓ, the value of f is unchanged. Suppose then that the global optimum x^* over $\mathcal{D}$ has $x_1^* \ge \ell$ or $x_{n+1}^* \ge \ell$, and $x_i^* > \ell$ for some other component. By shrinking x_i^* to ℓ while stretching the larger of x_1^* or x_{n+1}^*, we can eventually transform x^* into a point in one of the four regions without changing its value under f. Thus the best of the minima of the four regions, call it $\widetilde{x}$, is not a global minimum over $\mathcal{D}$ only if for all global minima x^* over $\mathcal{D}$, $x_1^* < \ell$, $x_{n+1}^* < \ell$, and in some other component $x_i^* > \ell$. Shrinking x_i^* and stretching x_1^* or x_{n+1}^* as before shows that suboptimality of $\widetilde{x}$ is due only to error in $\widetilde{x}_1$ or $\widetilde{x}_{n+1}$. However, as there are no linking clones by which to estimate $\widetilde{x}_1$ and $\widetilde{x}_{n+1}$, the hybridization data provides no direct information by which to reconstruct these two exterior distances, and their estimates should be regarded with suspicion in any reconstruction. Thus, if the biologist interprets the output $\widetilde{x}$ with the understanding that when $\widetilde{x}_i = \ell$ for some component, this distance may exceed ℓ in the true map, and that $\widetilde{x}_1$ and $\widetilde{x}_{n+1}$ may be inaccurate, then reporting the global optimum $\widetilde{x}$ of the four regions is reasonable.

4.2 Gradient descent

The *gradient* of f at point $\mathbf{p}$ is the vector

$$\operatorname{grad} f(\mathbf{p}) \quad := \quad \Big(\frac{\partial}{\partial x_1} f(\mathbf{p}), \; \cdots, \; \frac{\partial}{\partial x_n} f(\mathbf{p}) \Big),$$

where the kth component of the gradient is the partial derivative of f with respect to x_k evaluated at $\mathbf{p} = (p_1, \ldots, p_n)$,

$$\frac{\partial}{\partial x_k} f(\mathbf{p}) \quad = \quad \sum_{1 \le i \le m} \frac{1}{f_i(\mathbf{p})} \Big(b_{ik}^\pi \; u(p_k) \; - \; b_{in+1}^\pi \; u(p_{n+1}) \Big),$$

where $u(\cdot)$ is the unit step function defined before and p_{n+1} is defined in the same way as x_{n+1}. A basic fact in multivariable calculus is that the direction of greatest decrease of f at $\mathbf{p}$ is $\mathbf{v} = -\operatorname{grad} f(\mathbf{p})$ [17].

The procedure known as *gradient descent* [23] starts from a point $\mathbf{p}$, computes the negative gradient direction $\mathbf{v}$ at $\mathbf{p}$, moves to the point $\mathbf{p}'$ that minimizes f along the ray $\mathbf{p}+s\mathbf{v}$, and repeats, stopping once a point is reached at which the gradient vanishes. In the unconstrained problem of minimizing f over $\mathcal{R}^n$, such a point is a local minimum, and since f is convex, when gradient descent halts it has found a global minimum of the unconstrained problem.

For the constrained problem, however, of minimizing f over a region $\mathcal{C} \subseteq \mathcal{R}^n$, the negative gradient direction $\mathbf{v}$ at a point $\mathbf{p}$ on the boundary of $\mathcal{C}$ may be directed outside $\mathcal{C}$, in which case we cannot move along $\mathbf{v}$, yet another direction $\mathbf{v}'$ at $\mathbf{p}$ that is directed inside $\mathcal{C}$ may exist along which f decreases, albeit at a slower rate. Let us call a direction $\mathbf{v}$ at a point $\mathbf{p} \in \mathcal{C}$ *feasible* if it is possible move along $\mathbf{v}$ from $\mathbf{p}$ and remain in $\mathcal{C}$. In general, the feasible direction $\mathbf{v}$ of greatest decrease in f at a point $\mathbf{p}$ can be found as follows.

The boundaries of a region $\mathcal{C} = \mathcal{D}_{ab}$ are given by constraints that are hyperplanes. At point $\mathbf{p} \in \mathcal{C}$, compute the negative gradient direction $\mathbf{v} = -\operatorname{grad} f(\mathbf{p})$, and determine which of the bounding hyperplanes are tight. Let the list of tight hyperplanes for which $\mathbf{v}$ points outside the halfspace given by the hyperplane be H_1, ..., H_k. Take $\mathbf{v}^{(0)} = \mathbf{v}$, and successively project $\mathbf{v}^{(0)}$ onto H_1 to obtain $\mathbf{v}^{(1)}$, then project $\mathbf{v}^{(1)}$ onto H_2 to obtain $\mathbf{v}^{(2)}$, and so on. The vector $\mathbf{v}^{(k)}$ resulting from the final projection onto H_k is the feasible direction of greatest decrease at $\mathbf{p}$. If $\mathbf{v}^{(k)} = \mathbf{0}$, then $\mathbf{p}$ is a local minimum of f in $\mathcal{C}$.

Given the feasible direction $\mathbf{v}$ of greatest decrease, we compute the largest value $t > 0$ for which $\mathbf{p} + t\mathbf{v} \in \mathcal{C}$. As f is convex, the one-dimensional problem of minimizing f along $\mathbf{p} + s\mathbf{v}$ for $s \in [0, t]$ can be solved by a form of binary search known as *bisection* [23].

This completes the description of our approach to evaluating $g(\pi)$. Over each of the four regions $\mathcal{D}_{+1+1}, \ldots, \mathcal{D}_{-1-1}$, we compute a global minimum by constrained gradient descent using bisection, and take the best of the four minima. Computing the gradient at a given point takes time $\Theta(mn)$, which dominates the time to find the best feasible direction by successive projection, and is also the time to compute derivatives at each step during bisection. As reaching a local minimum can involve several gradient descent iterations, and each iteration can involve several bisection steps, the entire procedure is expensive. To find a good π we use the local-search heuristic known as simulated annealing, calling the above procedure to evaluate $g(\pi)$ on each candidate probe order.

5 Preliminary results

We now present some very preliminary results with an implementation of this approach written by the second author.

In the first experiments we ran the implementation on simulated data. For our parameters we picked values identical to those for chromosome VI of the fungus *Aspergillus nidulans*, which has been mapped using the sampling-without-replacement protocol [22]. This involved $m = 1118$ clones, $n = 77$ probes, a clone length of $\ell = 40$ kb, and a chromosome length of $L = 3500$ kb, which corresponds to a coverage of nearly 13. A false positive and false negative probability of $\rho = \eta = 0.2\%$ were used, which are the estimated error rates for the mapping project. Clones were thrown at random across the chromosome with the uniform distribution, a probe set of nonoverlapping clones was chosen, and the corresponding hybridization matrix H with false positives and false negatives was generated.

We first tested how well the approach recovered the true spacing, which was known for the simulated data, by running the constrained gradient descent procedure with the true probe order π. This is summarized in Table 1 for the gradient descent started from a completely uniform initial spacing, and an initial spacing obtained by a linear programming approximation (which will be described in the full paper). The hope was that a more sophisticated method for choosing an initial spacing would lead to faster convergence to a local minimum. As Table 1 shows, this was not the case. Starting from a uniform spacing took fewer iterations of gradient descent, and fewer total bisection steps. It is interesting that both approaches found a final spacing with better likelihood than the true spacing, which had a value of 6649.32.

As a measure of the error between the true spacing and the computed spacings, we used the root-mean-square error (RMS). Interestingly, the linear programming spacing had greater initial error because the two exterior distances x_1 and x_{n+1} were not well-estimated from the hybridization data, and the uniform spacing happened to give better estimates for these exterior distances. The computation time using either initial spacing was around 5 minutes on a Sun UltraSPARC 1 with a 167 MHz chip. The final RMS error of 3.7 kb is roughly 9% of the clone length.

Clearly there is a limit to the accuracy to which one can recover the true spacing from the discrete data of a hybridization matrix, which is essentially giving counts of linking clones. We can show that every method of recovering spacings must in the worst case have a root-mean-square error of at least $\epsilon = \frac{2}{m}(L - \ell)$. For the above data, $\epsilon \approx 6.2$ kb. In comparison, the final error in Table 1 is around 60% of this worst-case lower bound.

Next we tested how well the simulated annealing approach combined with this procedure for evaluating f recovered the true probe order. We started from an initial π obtained by a greedy heuristic for the Hamming-distance traveling salesman objective. This initial π had 6 breakpoints with respect to the true π, and an initial likelihood of 6728.45. After about 12 hours on the above machine the simulated annealing procedure halted with a final π equal to the true order, with a final likelihood of 6470.52.

In the second experiments we ran the implementation on real mapping data from chromosome VI of *Aspergillus nidulans*, which took around 12 hours on the above machine. The computed probe order had 36 breakpoints with respect to the published order [22], which was obtained using simulated annealing on the Hamming-distance traveling salesman objective [25]. While our computed order clearly had little in common with the published order, for this mapping data the true order is not known.

6 Conclusion

We have presented a new maximum-likelihood approach for reconstructing the distances between probes for physical maps constructed by hybridizing equal-sized clones against a nonoverlapping clone-subset. This protocol has been used to successfully map several organisms, and yields a model whose likelihood function is sufficiently simple to permit a closed-form expression. The resulting formulation gives to our knowledge the first practical method for physical mapping from hybridization data that can reconstruct globally-optimal maximum-likelihood distances along maps.

Table 1 Recovering the spacing on data simulating chromosome VI of *Aspergillus nidulans*.

	LP-based initial spacing	Uniform initial spacing
Bisection steps	185	177
Gradient descent iterations	149	101
Initial RMS error	6.65 kb	5.66 kb
Final RMS error	3.78 kb	3.69 kb
Final likelihood	6610.41	6610.53

Further research Finding a provably optimal π under the objective $g(\pi) = \min_{x \in \mathcal{D}} f(\pi, x)$ appears formidable given that f is nonlinear, while attempting to find a good π through simulated annealing started from a random π appears slow given that $g(\pi)$ is expensive to evaluate. The following two-stage approach may be effective, however:

(1) Use a combinatorial approach with guaranteed performance to find an initial $\widetilde{\pi}$ that optimizes a simpler linear combinatorial objective $\widetilde{g}(\pi)$.

(2) Polish $\widetilde{\pi}$ under the original nonlinear objective g by local search to obtain a final π^* and spacing x^*.

For example, $\widetilde{g}$ could be the combinatorial 2-consecutive-ones objective of Christof and Kececioglu [8], which corresponds to the same likelihood model but without probe spacings. In fact, if $\widetilde{g}$ is sufficiently accurate to recover an acceptable $\widetilde{\pi}$, one might use the original objective f to simply recover the best spacing for $\widetilde{\pi}$. We suspect that the full f is not needed to recover the true probe order in practice, and that the real utility of our likelihood function f will be to infer probe spacings for probe orders computed by combinatorial methods.

The numerical techniques we used to compute $x^* \in \mathrm{argmin}_{x \in \mathcal{D}} f(\pi, x)$, namely gradient descent with bisection, are elementary, and it would be interesting to investigate whether convergence to x^* can be sped up by more sophisticated numerical techniques.

In taking $f(\pi, x)$ as our objective, which is equivalent to maximizing $p(H \mid \pi, x)$, not $p(\pi, x \mid H)$, we are implicitly assuming that the *a priori* probability density function on probe spacings, $p(x)$, is uniform. Unfortunately, even when the distribution of the left ends of clones is uniform, the density function on probe spacings is not. It would be interesting to work out the *a priori* probe spacing density function under a natural model of clone placement (which appears to be involved), and investigate whether its inclusion in the likelihood objective improves recovery of the true spacing.

Finally, a significant source of error not considered in our model is repeated DNA. When the chromosome contains a repeat R that happens to occur at the end of a probe P, the probe will have a false-positive hybridization with every clone that does not overlap P but contains the same repeat R. Examination of the hybridization matrix for chromosome VI of *Aspergillus nidulans* shows that the false positives do not appear to occur completely independently across the matrix, but appear to occur more frequently in certain columns. This suggests that repeats may be present. How to best incorporate repeats into the maximum likelihood objective is an interesting open problem, as it is not clear how to appropriately model both the number of repeat families and the number of copies in a family.

References

[1] Alizadeh, F., R.M. Karp, D.K. Weisser and G. Zweig. "Physical mapping of chromosomes using unique probes." Proceedings of the 5th ACM-SIAM *Symposium on Discrete Algorithms*, 489–500, 1994.

[2] Alizadeh, F., R.M. Karp, L.A. Newberg and D.K. Weisser. "Physical mapping of chromosomes: A combinatorial problem in molecular biology." *Algorithmica* 13:1/2, 52–76, 1995.

[3] Arnold, J. and M.T. Cushion. "Constructing a physical map of the *Pneumocystis* genome." *Journal of Eukaryotic Microbiology* 44, 8S, 1997.

[4] Arratia, R., E.S. Lander, S. Tavare and M.S. Waterman. "Genomic mapping by anchoring random clones: A mathematical analysis." *Genomics* 11, 806–827, 1991.

[5] Ben-Dor, A. and B. Chor. "On constructing radiation hybrid maps." Proceedings of the 1st ACM *Conference on Computational Molecular Biology*, 17–26, 1997.

[6] Chor, B. and M. Sudan. "A geometric approach to betweenness." Proceedings of the *European Symposium on Algorithms*, Springer-Verlag Lecture Notes in Computer Science 979, 227–237, 1995.

[7] Christof, T., M. Jünger, J. Kececioglu, P. Mutzel, and G. Reinelt. "A branch-and-cut approach to physical mapping of chromosomes by unique end-probes." *Journal of Computational Biology* 4:4, 433–447, 1997.

[8] Christof, T. and J. Kececioglu. "Computing physical maps of chromosomes with nonoverlapping probes by branch-and-cut." Proceedings of the 3rd ACM *Conference on Computational Molecular Biology*, 115–123, 1999.

[9] Cuticchia, A.J., J. Arnold and W.E. Timberlake. "The use of simulated annealing in chromosome reconstruction experiments based on binary scoring." *Genetics* 132, 591–601, 1992.

[10] Fasulo, D.P., T. Jiang, R.M. Karp, R. Settergren, E.C. Thayer. "An algorithmic approach to multiple complete digest mapping." Proceedings of the 1st ACM *Conference on Computational Molecular Biology*, 118–127, 1997.

[11] Greenberg, D.S. and S. Istrail. "Physical mapping by STS hybridization: Algorithmic strategies and the challenge of software evaluation." *Journal of Computational Biology* 2:2, 219–273, 1995.

[12] Jain, M. and E.W. Myers. "Algorithms for computing and integrating physical maps using unique probes." Proceedings of the 1st ACM *Conference on Computational Molecular Biology*, 151–161, 1997.

[13] Jiang, T. and R.M. Karp. "Mapping clones with a given ordering or interleaving." Proceedings of the 8th ACM-SIAM *Symposium on Discrete Algorithms*, 400–409, 1997.

[14] Karp, R.M. and R. Shamir. "Algorithms for optical mapping." Proceedings of the 2nd ACM *Conference on Computational Molecular Biology*, 117–124, 1998.

[15] Lander, E.S. and M.S. Waterman. "Genomic mapping by fingerprinting random clones: A mathematical analysis." *Genomics* 2, 231–239, 1988.

[16] Lee, J.K., V. Dancik and M.S. Waterman. "Estimation for restriction sites observed by optical mapping using reversible-jump Markov chain Monte Carlo." Proceedings of the 2nd ACM *Conference on Computational Molecular Biology*, 147–152, 1998.

[17] Lengauer, T. *Combinatorial Algorithms for Integrated Circuit Layout.* John Wiley and Sons, Chichester, 1990.

[18] Mayraz, G. and R. Shamir. "Construction of physical maps from oligonucleotide fingerprint data." Proceedings of the 3rd ACM *Conference on Computational Molecular Biology*, 268–277, 1999.

[19] Mizukami, T., W.I. Chang, I. Garkatseve, N. Kaplan, D. Lombardi, T. Matsumoto, O. Niwa, A. Kounosu, M. Yanagida, T.G. Marr and D. Beach. "A 13 kb resolution cosmid map of the 14 Mb fission yeast genome by nonrandom sequence-tagged site mapping." *Cell* 73, 121–132, 1993.

[20] Mott, R., A. Grigoriev, E. Maier, J. Hoheisel and H. Lehrach. "Algorithms and software tools for ordering clone libraries: Application to the mapping of the genome *Schizosaccharomyces pombe*." *Nucleic Acids Research* 21:8, 1965–1974, 1993.

[21] Muthukrishan, S. and L. Parida. "Towards constructing physical maps by optical mapping: An effective, simple, combinatorial approach." Proceedings of the 1st ACM *Conference on Computational Molecular Biology*, 209–219, 1997.

[22] Prade, R.A., J. Griffith, K. Kochut, J. Arnold and W.E. Timberlake. "*In vitro* reconstruction of the *Aspergillus nidulans* genome." *Proceedings of the National Academy of Science USA* 94, 14564–14569, 1997.

[23] Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery. *Numerical Recipes in C.* Cambridge University Press, New York, 1992.

[24] Slonim, D., L. Kruglyak, L. Stein, and E. Lander. "Building human genome maps with radiation hybrids." Proceedings of the 1st ACM *Conference on Computational Molecular Biology*, 277–286, 1997.

[25] Wang, Y., R.A. Prade, J. Griffith, W.E. Timberlake and J. Arnold. "A fast random cost algorithm for physical mapping." *Proceedings of the National Academy of Science USA* 91, 11094–11098, 1994.

[26] Wilson, D.B., D.S. Greenberg and C.A. Phillips. "Beyond islands: Runs in clone-probe matrices." Proceedings of the 1st ACM *Conference on Computational Molecular Biology*, 320–329, 1997.

[27] Xiong, M., H.J. Chen, R.A. Prade, Y. Wang, J. Griffith, W.E. Timberlake and J. Arnold. "On the consistency of a physical mapping method to reconstruct a chromosome *in vitro*." *Genetics* 142, 267–284, 1996.

[28] Zhang, M.Q. and T.G. Marr. "Genome mapping by nonrandom anchoring: A discrete theoretical analysis." *Proceedings of the National Academy of Science USA* 90, 600–604, 1993.

A Deriving the objective function

In this appendix we derive the objective function $f(\pi, x)$. We first work out the probability of each type of overlap event, conditioned on a given probe order π and probe spacing x.

To simplify the notation, let

$$\begin{aligned}
\overline{\rho} &:= 1 - \rho, \\
\overline{\eta} &:= 1 - \eta, \\
\alpha &:= \overline{\eta}/\rho, \\
\beta &:= \eta/\overline{\rho}, \\
k_i &:= \sum_{1 \le j \le n} h_{ij}, \\
\overline{k}_i &:= n - k_i, \\
\chi_{ij}^{\pi} &:= h_{i\pi_j}, \\
\overline{\chi}_{ij}^{\pi} &:= 1 - h_{i\pi_j}, \\
G &:= L - \ell, \\
g &:= L - n\ell.
\end{aligned}$$

Then

- $\overline{\rho}$ is the probability of a true negative,
- $\overline{\eta}$ is the probability of a true positive,
- k_i is the number of ones in row i,
- $\overline{k}_i$ is the number of zeros in row i,
- G is the effective genome length, and
- g is the total gap length.

Assuming the left ends of clones are uniformly distributed across the chromosome, it suffices to determine the length of the interval corresponding to B_j^{π}, O_j^{π}, and N_j^{π}. Examining Figure 3,

$$\begin{aligned}
p(C_i \in N_j^{\pi} \mid \pi, x) &= \big(x_j - \min\{x_j, \ell\}\big)\big/G, \\
p(C_i \in O_j^{\pi} \mid \pi, x) &= \big(\min\{x_j, \ell\} + \min\{x_{j+1}, \ell\}\big)\big/G, \\
p(C_i \in B_j^{\pi} \mid \pi, x) &= \big(\ell - \min\{x_{j+1}, \ell\}\big)\big/G,
\end{aligned}$$

where the first equation holds for $1 \le j \le n+1$, the second equation holds for $1 \le j \le n$, and the third equation holds for $1 \le j < n$.

For a given clone C_i, we cannot know which overlap event occurred, as we only observe the hybridization results of C_i versus the probes given by row i of H (which contains errors). We can work out the probability of observing row i, however, conditioned on π, x, and the occurrence of a given overlap event.

Suppose event $C_i \in N_j^{\pi}$ occurred, where $1 \le j \le n+1$. In the absence of errors, row i of H should contain all zeros. Hence the observed row i contains

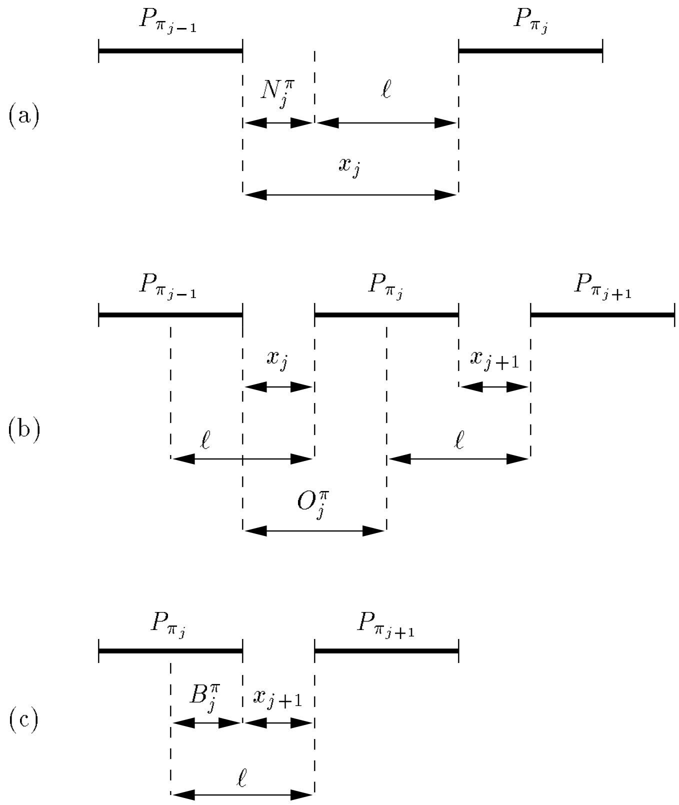

Figure 3 Lengths of regions corresponding to events (a) N_j^π, (b) O_j^π, and (c) B_j^π.

- no true positives,
- k_i false positives,
- $\overline{k}_i$ true negatives, and
- no false negatives.

Thus

$$p(H_i \mid \pi, x, C_i \in N_j^\pi) \;=\; \rho^{k_i}\, \overline{\rho}^{\overline{k}_i}.$$

Now suppose event $C_i \in O_j^\pi$ occurred, where $1 \le j \le n$. In the absence of errors, row i of H should contain a one at column π_j and a zero everywhere else. Hence the observed row i contains

- χ_{ij}^π true positives,
- $k_i - \chi_{ij}^\pi$ false positives,
- $\overline{k}_i - \overline{\chi}_{ij}^\pi$ true negatives, and
- $\overline{\chi}_{ij}^\pi$ false negatives.

Thus

$$p(H_i \mid \pi, x, C_i \in O_j^\pi) \;=\; \rho^{k_i}\, \overline{\rho}^{\overline{k}_i}\, \alpha^{\chi_{ij}^\pi}\, \beta^{\overline{\chi}_{ij}^\pi}.$$

Finally, suppose event $C_i \in B_j^\pi$ occurred, where $1 \le j < n$. In the absence of errors, row i of H should contain a one at column π_j, a one at column π_{j+1}, and a zero everywhere else. Hence the observed row i contains

- $\chi_{ij}^\pi + \chi_{ij+1}^\pi$ true positives,
- $k_i - (\chi_{ij}^\pi + \chi_{ij+1}^\pi)$ false positives,
- $\overline{k}_i - (\overline{\chi}_{ij}^\pi + \overline{\chi}_{ij+1}^\pi)$ true negatives, and
- $\overline{\chi}_{ij}^\pi + \overline{\chi}_{ij+1}^\pi$ false negatives.

Thus

$$p(H_i \mid \pi, x, C_i \in B_j^\pi) \;=\; \rho^{k_i}\, \overline{\rho}^{\overline{k}_i}\, \alpha^{\chi_{ij}^\pi + \chi_{ij+1}^\pi}\, \beta^{\overline{\chi}_{ij}^\pi + \overline{\chi}_{ij+1}^\pi}.$$

Let the aggregate events N^π, O^π, and B^π be

$$\begin{aligned} N^\pi &:= \bigcup_{1 \le j \le n+1} N_j^\pi, \\ O^\pi &:= \bigcup_{1 \le j \le n} O_j^\pi, \\ B^\pi &:= \bigcup_{1 \le j < n} B_j^\pi. \end{aligned}$$

Then since events N_j^π, O_j^π, and B_j^π are disjoint for all j,

$$\begin{aligned} &p(H_i \mid \pi, x, C_i \in N^\pi) \\ &\quad = \sum_{1 \le j \le n+1} p(H_i \mid \pi, x, C_i \in N_j^\pi)\; p(C_i \in N_j^\pi \mid \pi, x) \\ &\quad = \frac{\rho^{k_i}\, \overline{\rho}^{\overline{k}_i}}{G} \Big(g \;-\; \sum_{1 \le j \le n+1} \min\{x_j, \ell\} \Big), \\ &p(H_i \mid \pi, x, C_i \in O^\pi) \\ &\quad = \sum_{1 \le j \le n} p(H_i \mid \pi, x, C_i \in O_j^\pi)\; p(C_i \in O_j^\pi \mid \pi, x) \\ &\quad = \frac{\rho^{k_i}\, \overline{\rho}^{\overline{k}_i}}{G} \sum_{1 \le j \le n} \alpha^{\chi_{ij}^\pi}\, \beta^{\overline{\chi}_{ij}^\pi} \Big(\min\{x_j, \ell\} \;+\; \min\{x_{j+1}, \ell\} \Big), \\ &p(H_i \mid \pi, x, C_i \in B^\pi) \\ &\quad = \sum_{1 \le j < n} p(H_i \mid \pi, x, C_i \in B_j^\pi)\; p(C_i \in B_j^\pi \mid \pi, x) \\ &\quad = \frac{\rho^{k_i}\, \overline{\rho}^{\overline{k}_i}}{G} \sum_{1 \le j < n} \alpha^{\chi_{ij}^\pi + \chi_{ij+1}^\pi}\, \beta^{\overline{\chi}_{ij}^\pi + \overline{\chi}_{ij+1}^\pi} \Big(\ell - \min\{x_{j+1}, \ell\} \Big). \end{aligned}$$

Finally, since the sample space for C_i conditioned on π and x is the disjoint union of N^π, O^π, and B^π,

$$\begin{aligned} p(H_i \mid \pi, x) \;&=\; p(H_i \mid \pi, x, C_i \in N^\pi) \;+ \\ &\;\; p(H_i \mid \pi, x, C_i \in O^\pi) \;+ \\ &\;\; p(H_i \mid \pi, x, C_i \in B^\pi) \\ &=\; c_i \Big(a_i^\pi \;-\; \sum_{1 \le j \le n+1} b_{ij}^\pi \, \min\{x_j, \ell\} \Big), \end{aligned}$$

where

$$a_i^\pi \;:=\; g \;+\; \ell \sum_{1 < j \le n} d_{ij}^\pi\, d_{ij-1}^\pi, \tag{2}$$

$$b_{ij}^\pi \;:=\; (1 - d_{ij}^\pi)(1 - d_{ij-1}^\pi), \tag{3}$$

$$c_i \;:=\; \rho^{k_i}\, \overline{\rho}^{\overline{k}_i} \big/ G, \tag{4}$$

$$d_{ij}^\pi \;:=\; \begin{cases} \alpha^{\chi_{ij}^\pi}\, \beta^{\overline{\chi}_{ij}^\pi}, & 1 \le j \le n; \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

For a given probe order π, finding a spacing vector x that maximizes $p(H \mid \pi, x)$ is equivalent to computing

$$\begin{aligned} &\operatorname*{argmax}_{x \in \mathcal{D}} \ln p(H \mid \pi, x) \\ &\quad = \operatorname*{argmin}_{x \in \mathcal{D}} \; -\ln p(H \mid \pi, x) \\ &\quad = \operatorname*{argmin}_{x \in \mathcal{D}} \; -\ln \prod_{1 \le i \le m} p(H_i \mid \pi, x) \\ &\quad = \operatorname*{argmin}_{x \in \mathcal{D}} \; -\sum_{1 \le i \le m} \ln p(H_i \mid \pi, x) \\ &\quad = \operatorname*{argmin}_{x \in \mathcal{D}} \; -\sum_{1 \le i \le m} \Big(\ln c_i \;+ \\ &\qquad\qquad \ln \Big(a_i^\pi \;-\; \sum_{1 \le j \le n+1} b_{ij}^\pi \, \min\{x_j, \ell\} \Big) \Big) \\ &\quad = \operatorname*{argmin}_{x \in \mathcal{D}} \; -\sum_{1 \le i \le m} \ln \Big(a_i^\pi \;-\; \sum_{1 \le j \le n+1} b_{ij}^\pi \, \min\{x_j, \ell\} \Big). \end{aligned}$$

The right-hand side is our objective function $f(\pi, x)$.

Note that the coefficients a_i^π and b_{ij}^π depend only on π and H, and are constants with respect to the x_j.