Computing physical maps of chromosomes with nonoverlapping probes by branch-and-cut

Thomas Christof* John Kececioglu†

Abstract

We introduce a new combinatorial formulation of chromosome physical-mapping by the 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 subset of the clones called *probes*, which are designed to form a maximal nonoverlapping clone-subset. The output of the protocol is the clone-probe hybridization matrix H. The problem of finding a maximum-likelihood reconstruction of the order of the probes along the chromosome in the presence of false positive and negative hybridization error is equivalent to finding the minimum number of entries of H to change to zeros so that the resulting matrix has at most 2 ones per row, and the consecutive-ones property across rows. This combinatorial problem, which we call *2-Consecutive-Ones Mapping*, has a concise integer linear-programming formulation, to which we apply techniques from polyhedral combinatorics. The formulation is unique in that it does not explicitly represent the probe permutation, and in contrast to prior linear-programming approaches, the number of variables is small: in practice, linear in the number of clones. We derive a large class of facet-defining inequalities for the 2-consecutive-ones polytope that we call the augmented k-degree inequalities, and we show that the basic k-degree class can be efficiently separated using bipartite matchings. Computational results with an implementation of the resulting branch-and-cut algorithm applied to both simulated and real data from the complete genome of *Aspergillus nidulans* show that we can solve many problems to provable optimality and find maps of higher quality than previously possible.

*Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany. Email: thomas.christof@iwr.uni-heidelberg.de. Research performed during a DAAD Postdoctoral Fellowship at the University of Georgia.

†Corresponding author. Department of Computer Science, University of Georgia, Athens, GA 30602-7404, USA. Email: kece@cs.uga.edu. Research supported by a National Science Foundation CAREER Award, Grant DBI-9722339.

Keywords Computational biology, physical mapping of chromosomes, sampling without replacement protocol, consecutive ones property, branch-and-cut

1 Introduction

Physical mapping of chromosomes is a basic task in molecular biology that is essentially the problem of reconstructing the order and location of features of biological interest along a chromosome from indirect measurements. A wide variety of approaches have been developed depending on the type of data that is collected, including mapping from clone-probe hybridization data with unique [1, 15, 9, 17] or nonunique probes [2], mapping from restriction-fragment length data [18, 13], radiation-hybrid mapping [4, 31], and optical mapping [26, 21, 29, 23], and a variety of computational techniques have been suggested, including simulated annealing [12, 25, 14, 33, 1, 2], linear programming [9, 17, 10], and semidefinite programming [7].

In this paper we study a new formulation of physical mapping by a procedure known as the sampling-without-replacement protocol. The protocol is inexpensive, easy to perform, and has been used to map several organisms, including *Schizosaccharomyces pombe* [24], *Aspergillus nidulans* [30], and *Pneumocystis carinii* [3]. Projects to map *Neurospora crassa* and *Aspergillus flavus* by the protocol are currently underway.

In the protocol, a library $\mathcal{L}$ of many overlapping clones, each of which samples an interval of the chromosome, are screened to have equal length and arrayed on a plate. Then a subset $\mathcal{P} \subseteq \mathcal{L}$ of the clones is chosen as follows. Initially let $\mathcal{P} = \emptyset$ and $\mathcal{S} = \mathcal{L}$. At step i, a clone $P_i \in \mathcal{S}$ is randomly chosen, removed from $\mathcal{S}$, and added to $\mathcal{P}$. Clone P_i is then hybridized against the library by extracting complementary DNA from both of its ends, washing the DNA over the arrayed plate, and recording the clones to which the DNA sticks. All clones with a positive hybridization result to P_i are removed from $\mathcal{S}$. The process is then repeated until $\mathcal{S} = \emptyset$.

We call the final set $\mathcal{P}$ the *probe set* and the set $\mathcal{C} = \mathcal{L} - \mathcal{P}$ the *clone set*. The output of the protocol is the clone-probe hybridization matrix H, which records the outcomes of all hybridizations between the probes in $\mathcal{P}$ and the clones in $\mathcal{C}$.

Note that in this protocol, probes correspond to intervals of the chromosome, as opposed to sequence-tagged-site mapping, for example, in which probes correspond to points along the chromosome. Note also that if a clone overlaps with a probe P_i, then since all clones have the same length, its interval must contain at least one end of P_i. Thus in the absence of false negative error, the clone will hybridize with P_i and be removed from $\mathcal{S}$ at the ith step. This implies that in the absence of error the probe set $\mathcal{P}$ generated by the above process is a maximal nonoverlapping subset of the clones.

The main goal of physical mapping by this protocol is to recover the order of the probes across the chromosome from hybridization matrix H. Recovering this order is very useful. Suppose that for every pair of probes (P, Q) that are consecutive in the order we can identify from matrix H a clone that overlaps with both P and Q, called a *linking clone*. Then probe set $\mathcal{P}$ together with a linking clone for each consecutive pair is a minimal set of clones that covers the chromosome, called a *tiling set*. By sequencing just the probes and clones in a tiling set and overlapping their sequences in the appropriate order, we can reconstruct the DNA sequence for the chromosome in a cost-effective manner.

Related work Prior work on mapping by the sampling-without-replacement protocol, by Cuticchia, Arnold and Timberlake [12], Wang, Prade, Griffith, Timberlake and Arnold [33], and Mott, Grigoriev, Maier, Hoheisel and Lehrach [25], 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 [34] 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.

In this paper we study the problem of finding a *maximum-likelihood* reconstruction of the order of the probes along the chromosome from hybridization matrix H in the presence of false positive and false negative error. The problem of finding a maximum-likelihood reconstruction of the *distance* between probes, as well as their order, is studied in [22].

Plan of the paper In the following section we show that finding a maximum-likelihood probe order is equivalent to a combinatorial problem that we call 2-*Consecutive-Ones Mapping*. Section 3 gives an integer linear-programming formulation of this problem. Section 4 examines the structure of the polytope of feasible solutions to the integer program, and characterizes a general class of facet-defining inequalities that we call the augmented k-degree inequalities. Section 5 shows that the separation problem for the basic k-degree class can be efficiently solved using maximum-weight bipartite matchings. Solving the separation problem identifies cutting planes, which are inequalities in the class that are violated by a fractional solution to the current linear-programming relaxation of the integer program. Adding these cutting planes to the linear program yields a tighter relaxation. Section 6 describes how feasible solutions are obtained from the fractional solutions generated after solving successive linear programs. Section 7 presents computational results from experiments with an implementation of this branch-and-cut approach to 2-Consecutive-Ones Mapping on simulated data and real data from all eight chromosomes of the fungus *Aspergillus nidulans*. We then close with some directions for further work.

2 The problem

In our maximum-likelihood formulation of the problem, we do not model the sequential process of generating probe set $\mathcal{P}$. Hence we assume that $\mathcal{P}$ is a pairwise-nonoverlapping subset of the clone library $\mathcal{L}$. We also do not model error due to chimeric clones, which are clones that sample more than one interval of the chromosome, error due to deletions, which are segments of clones that get spliced out during cloning, or error due to repeated DNA. Clones for the mapping projects using this protocol at the University of Georgia are produced by cosmids, which are short enough that chimerism and deletions are not a significant problem. False hybridizations due to repeats are treated as a series of independent false positives.

We model the problem of reconstructing the probe order from the hybridization matrix in the presence of false positive and false negative hybridization error as follows. Let $\{C_1, \ldots, C_m\}$ be the set of m clones and $\{P_1, \ldots, P_n\}$ the set of n probes. The input to the problem is the $m \times n$ clone-probe hybridization matrix $H = (h_{ij})$, where

$$h_{ij} \quad = \quad \begin{cases} 1, & P_j \text{ hybridizesto } C_i; \\ 0, & \text{otherwise} \end{cases}$$

The output is a permutation π of the columns of H that maximizes $P(\pi \mid H)$, the probability that π is the true order of the probes across the chromosome, given that H is the observed hybridization matrix.

By Bayes' rule,

$$P(\pi \mid H) \quad = \quad \frac{P(H \mid \pi)\; P(\pi)}{P(H)},$$

where $P(H \mid \pi)$ is the probability of observing H given that π is the true order, $P(\pi)$ is the probability of order π occurring in nature, and $P(H)$ is the probability of observing H. Since names are randomly assigned to probes, $P(\pi) = 1/n!$ which is independent of π. The denominator $P(H)$ is independent of π as well. Thus both quantities are constants in the maximization, and it suffices to find a probe order π that maximizes $P(H \mid \pi)$.

Let n_ρ be the number of false positives in H given that π is the correct probe order, and n_η be the number of false negatives in H. Then assuming that the entries of H have been corrupted randomly and independently with false positive probability $\rho < 1/2$ and false negative probability $\eta < 1/2$, maximizing $P(\pi|H)$ is equivalent to minimizing the objective function

$$f(\pi) \quad := \quad c_\rho\, n_\rho \;+\; c_\eta\, n_\eta,$$

where $c_\rho = \ln \frac{1-\eta}{\rho}$ and $c_\eta = \ln \frac{1-\rho}{\eta}$.

Let H^* be matrix H with all hybridization errors corrected and its columns permuted by π. Since all clones have the same length, a clone can overlap with at most two consecutive probes. This implies that each row of H^* must have at most 2 ones, and that a row with 2 ones must have them in consecutive columns. An important consequence is that $n_\eta = 0$, since for any correction H^* that declares an entry to be a false negative, there is another correction that declares the entry to be a true negative and is more likely. In other words, the problem is equivalent to finding a correction with the minimum number of false positives. From a practitioner's point of view, it is not possible to recover false negative error from matrix H.

We say that a 0-1 matrix A has the *2-consecutive-ones property* if (1) every row of A contains at most 2 ones, and (2) there is a permutation of the columns of A so that in each row the ones are consecutive. We can now state our mapping problem as follows.

Definition 1 (2-Consecutive-Ones Mapping)
The *2-Consecutive-Ones Mapping Problem* is the following. Given a 0-1 hybridization matrix H with rows corresponding to clones and columns corresponding to probes, find the minimum number of entries of H to change to zeros so that the resulting matrix has the 2-consecutive-ones property. □

Given a solution to 2-Consecutive-Ones Mapping, it is easy to recover an optimal order of the probes. After correcting the false positives identified by the solution, construct a graph G with a vertex for each column and an edge between two vertices if some row has ones in both columns. For every solution, graph G is a collection of vertex-disjoint paths. Concatenating the paths in any order yields an optimal probe order π.

We now express this problem in the language of integer programming.

3 An integer programming formulation

Given hybridization matrix $H = (h_{ij})$, we construct an integer linear program as follows. The variables x_{ij} correspond to all entries of H with $h_{ij} = 1$. The trivial inequalities are given by the positivity contraints,

$$0 \le x_{ij} \le 1, \quad \text{for all } i, j, \tag{1}$$

and the row-sum constraints,

$$\sum_j x_{ij} \le 2, \quad \text{for all } i. \tag{2}$$

The remaining inequalities are in two classes. The *k-degree inequalities*, for $3 \le k \le \min\{m, n-1\}$, have the form,

$$\begin{pmatrix} 1 & 1 & & & \\ 1 & & 1 & & \\ \vdots & & & \ddots & \\ 1 & & & & 1 \end{pmatrix} x \quad \le \quad k+2. \tag{3}$$

This notation means the following. The $k \times (k+1)$, 0-1 matrix of coefficients on the left-hand side is a column of k ones joined to a $k \times k$ identity matrix. The above inequality denotes that for any choice of $k+1$ columns of H and k rows of H, the inner product of the above coefficients with the variables x_{ij} corresponding to the entries at these rows and columns must have value at most $k+2$. Consider the graph G referred to after the problem definition that has a vertex for each column and an edge between two vertices if in the correction of H there is a row with ones in both columns. The above inequalities enforce that the degree of every vertex of G is at most 2.

The *k-cycle inequalities*, for $3 \le k \le \min\{m, n\}$, have the form,

$$\begin{pmatrix} 1 & 1 & & & & \\ & 1 & 1 & & & \\ & & 1 & 1 & & \\ & & & & \ddots & \\ 1 & & & & & 1 \end{pmatrix} x \quad \le \quad 2k-1. \tag{4}$$

The $k \times k$, 0-1 matrix of coefficients on the left consists of k cyclic shifts of the row vector $1100\cdots0$. Again the inequality denotes that for any choice of k columns and k rows of H, the inner product of these coefficients with the corresponding variables at these rows and columns must have value at most $2k-1$. In graph G referred to above, these inequalities enforce that G contains no cycles.

The objective is to minimize the linear function

$$\sum_{ij} (1 - x_{ij}),$$

over all integral x_{ij} that satisfy inequalities (1) through (4). By the arguments concerning graph G given after the problem definition, this finds a correction of H into a matrix with the 2-consecutive-ones property that minimizes the number of false positives.

Notice that the number of variables is equal to the number of ones in matrix H. In practice H is extremely sparse, with very few rows containing more than 5 ones; for such data, the number of variables is essentially linear in the number of clones m.

4 The 2-consecutive-ones polytope

The integer programming formulation gives only a first, weak relaxation of the 2-consecutive-ones problem. Essential for the success of the branch-and-cut algorithm is the knowledge of additional inequalities which are valid for the associated polytope that is spanned by all feasible solutions. We define this polytope as follows.

Let the set of all 0-1 matrices satisfying the consecutive-ones property (COP) be

$$\mathcal{X}_{COP}^{(m,n)} \;:=\; \{X \in \{0,1\}^{(m,n)} \mid X \text{ has the COP for rows}\}.$$

Then the *2-consecutive-ones polytope* for our integer program is the convex hull of the set of all such matrices with at most 2 ones per row:

$$P_{2COP}^{(m,n)} \;:=\; \text{conv}(\{X \mid X \in \mathcal{X}_{COP}^{(m,n)} \text{and} \sum_j x_{ij} \le 2, \text{for all } i\}).$$

In the Appendix we characterize some of the basic properties of this polytope. In addition, we present the facet-defining class of the augmented k-degree inequalities which generalize the k-degree inequalities.

5 Separation

Our branch-and-cut approach starts with the linear-programming relaxation consisting of the positivity inequalities and the row-sum inequalities, and repeatedly takes an optimal solution to the current relaxation, finds any inequalities in the polyhedral description that are violated by the fractional solution, and adds them to the current linear program, until the solution is integral and feasible for the integer program. If such an integral feasible solution is found, it is an optimal solution to the integer program. If no such solution is found after a fixed number of iterations, or no violated inequalities can be identified, the value of the final relaxation is a lower bound on the value of the minimum solution, and we can obtain a feasible solution by modifying the fractional variables as described in the next section.

The task of identifying violated inequalities is called *separation*. We observe that the most important inequalities in practice for obtaining a tight relaxation appear to be the augmented k-degree inequalities. Below we sketch how to exactly separate the basic k-degree inequalities, and how to heuristically separate the augmented inequalities using this exact separator.

Given a fractional solution x to the current linear-programming relaxation, we identify all violated k-degree inequalities as follows. We first consider all possibilities for the first column c of the coefficient matrix. (This is the column consisting of k ones.) For a given choice c for the column of H that corresponds to this column of the coefficient matrix, we construct a weighted bipartite graph G as follows. One half of the bipartition has n vertices corresponding to the columns of H, and the other half has m vertices corresponding to the rows of H. For every nonzero variable x_{ic}, we consider all the other nonzero variables x_{ij} in a column $j \neq c$. We then place an edge in G between the row-vertex corresponding to i and the column-vertex corresponding to j that has weight $x_{ic} + x_{ij}$.

It is not hard to show that every matching in G of cardinality k that has weight exceeding $k + 2$ corresponds to a violated k-degree inequality. By enumerating bipartite matchings of cardinality k in G in order of decreasing weight, we can identify as many violated k-degree inequalities as we like. Each successive violation can be output in $O(k(m + n \log n))$ time, and we can even output all violations in order of increasing k. This uses a well-known technique for generating next-best solutions to combinatorial optimization problems [16], specialized to weighted bipartite matchings [6].

The augmented k-degree inequalities are separated heuristically in two steps. An augmented k-degree inequality $f^T x \leq f_0$ (see Definition 3) can be seen as composed of two parts, q rows in which we have $f_{ic} = 2$ and p rows with $f_{ic} = 1$. The first step of the procedure identifies the rows and columns with $f_{ic} = 2$, in the second step the algorithm for separating the regular k-degree inequalities is used for identifying the rows and columns for which $f_{ic} = 1$. For the first step we choose a fixed number q, say $q \leq 5$, and use a template of the first part of the inequality (the rows with $f_{ic} = 2$). The number of ones in a row in the template is set to the maximal number of non-zero variables of the LP solution in a row. The template defines the structure of the first part of the inequality, but the column c and the involved rows and columns are not yet identified. Using the template of the inequality, in a local search procedure similar to the separation of small instance relaxation [8], now the rows and columns are assigned to the corresponding columns of the LP solution such that the scalar product is maximized. An assigment with value γ, greater than $2q$, gives a candidate for the first part of the inequality. Before we proceed with step 2, we reduce the columns that are involved in the inequality to c and the columns with nonzero values in the corresponding entries of the LP solution. In the second step, using the separation technique for separation of the k-degree inequalities on the system reduced by all rows and columns which are inolved in the first part of the inequality, we identify the second part of inequality. An inequality is violated if we find a matching of cardinality p with weight greater than $p + 2 + 2q - \gamma$.

If no violated augmented k-degree inequalities are found, we separate the cycle inequalities. It is easy to see that a violated cycle inequalities corresponds to an cycle with weight greater 1 in the bipartite row-column graph with edge weights $1 - x_{ij}$, where x is the LP solution.

6 Constructing a feasible solution

Given a fractional solution x to the linear-programming relaxation, we obtain a feasible solution for the integer program as follows.

We construct an edge-weighted graph G whose vertices correspond to the columns of H. Graph G has an edge between two vertices if in some row solution x has nonzero variables in both columns. The weight of an edge between vertices corresponding to columns j and k is $\sum_i (x_{ij} + x_{ik})$, the total weight between those two columns across all rows. We then sort the edges of G by decreasing weight, and scanning the sorted list, greedily form a collection of vertex-disjoint paths. Concatenating the paths gives a permutation π of the columns of H.

Given such a π, it is straightforward to identify the corresponding false positives in H. In every row with at least 2 ones, we search for a pair of consecutive ones. If such a pair exists, we change all other ones to zeros. If no such pair exists, we arbitrarily choose an isolated one, and again change all other ones to zeros.

This procedure is invoked to construct a feasible solution after solving each successive linear program. On termination, we output the best of these feasible solutions.

7 Computational results

We now summarize results from computational experiments on real and simulated data. Our implementation of the branch-and-cut code for the 2-COP problem uses ABACUS 2.2 [32, 20] with LP Solver Cplex. The Hamming-distance TSP's are solved by the branch-and-cut code for the traveling salesman problem of Jünger, Naddef, Reinelt and Thienel [19, 27, 28].

7.1 Real data

The first set of experiments tested our branch-and-cut approach on mapping data generated by the sampling-without-replacement protocol for the eight chromosomes of the fungus *Aspergillus nidulans* [30]. Each data set for a chromosome has roughly 100 probes and 1000 clones.

The results are summarized in Table 1. The second column is the number of false positives in the published map, obtained by simulated annealing applied to a Hamming-distance traveling-salesman problem [33]. The third column is the number of false positives in a map given by an *optimal* solution to the Hamming-distance TSP. The fourth and fifth columns are the number of false positives in the solution computed by our branch-and-cut code for 2-Consecutive-Ones Mapping (2-COP), after a computation time of 15 minutes and 24 hours respectively. The last two columns give a lower bound on the number of false positives in any map for the chromosome, again obtained by our branch-and-cut software after a computation time of 24 hours and 15 minutes respectively. These times are for our code running on a Sun Ultra 4 with a 296 MHz UltraSPARC-II chip.

Table 2 gives statistics on the amount of work performed by the branch-and-cut code on the *Aspergillus nidulans* data for the 15 minute runs. The second column gives the number of tree nodes explored in the branch-and-cut search tree, the third columns gives the total number of linear programs that were solved, the fourth column gives the total number of constraints that were added to these linear programs across all iterations, and the last column gives the number of variables in the linear programs. At each iteration, up to

Table 1: Number of false positives in reconstructed maps for *Aspergillus nidulans*.

Chromosome	Published	Hamming-TSP	2-COP		Lower bound	
			15 min	24 hr	24 hr	15 min
I	447	426	414	411	382	375
II	475	477	440	436	419	411
III	747	690	666	659	610	538
IV	363	327	306	304	295	290
V	506	497	475	474	448	433
VI	397	362	350	348	333	325
VII	566	521	507	502	475	459
VIII	631	606	585	576	540	498

Table 2: Effort expended in the branch-and-cut tree on real data.

Chromosome	Tree nodes	Linear programs	Constraints	Variables
I	1	390	194,500	908
II	1	428	213,500	919
III	1	390	194,500	1,285
IV	23	498	218,069	704
V	1	433	216,000	964
VI	1	552	275,500	756
VII	1	371	185,000	1,075
VIII	1	348	173,500	1,233

500 constraints were added to the linear program. (These are the violated inequalities found by the separation procedures.)

The maps computed by both the exact Hamming-TSP code and our 2-COP code differed markedly from the published maps. Figure 1 shows the 2-COP permutation plotted versus the published probe order for chromosome I. Unfortunately, for real data we cannot know the true order.

7.2 Simulated data

The next set of experiments tested the Hamming-TSP and 2-COP approaches on simulated data. The results are summarized in Table 3. Each row reports statistics averaged over 10 simulations. In all experiments, the number of clones is 1500, and the coverage is 10. Since probes are chosen as a non-overlapping subset of the clones, the number of probes is not constant, but varies between 95 and 110. The false negative probability in all experiments is 5%. The false positive probability varies and is given in the first column by ρ.

In these simulations we knew the true order of probes, and we measured the fraction of pairs of adjacent probes that were common to both the true and computed orders, which we call the *quality* of the reconstruction. Columns 2 and 3 give the average number of false positives and number of contigs in the true order. Columns 4 through 6 give the average quality, number of false positives, and number of contigs in the optimal Hamming-TSP order, which was found in around five minutes. Column 7 through 9 report these statistics for the 2-COP order, computed by limiting our code to 15 minutes of computation.

As can be seen, the quality of the 2-COP solution found by minimizing the number of false positives is consistently better than the quality of the optimal Hamming-TSP solution. With both approaches, the number of contigs in the reconstruction is much smaller than the true number.

It is interesting that the number of false positives of the Hammming-TSP is closer to the true number of errors than for the 2-COP solutions, while the quality of the 2-COP solutions is consistently better than the Hamming-TSP solutions. This suggests that for recovering the probe order, the 2-COP formulation may be a superior model for the sampling without replacement protocol than the Hamming-TSP model.

These experiments had more clones, but lower coverage, than the *Aspergillus nidulans* data. Table 4 summarizes results from additional experiments with 1000 clones and coverage of 10 and 15, with lower false positive rate and a false negative rate of 1%. The average number of probes is around 70. Each row again reports an average over 10 experiments. As can be seen, the quality of the solutions is much better, with the 2-COP quality still consistently better than the Hamming-TSP quality. The 2-COP code was limited to 30 minutes of computation on these experiments, in which time most of these problems were solved to optimality.

8 Conclusion

We have presented a new approach to computing physical maps constructed by hybrizing equal-sized clones against a nonoverlapping clone-subset called the probes. The approach finds a most-likely probe order in the presence of false positive and negative hybridization error by a technique known as branch-and-cut. Experiments with a preliminary implementation show that in a short amount of computing time the approach can find higher-quality maps than previously possible.

Further research The problem of reconstructing both the order of probes and their *spacing* across the chromosome from the clone-probe hybridization matrix has been recently studied by Kececioglu, Shete and Arnold [22]. Their

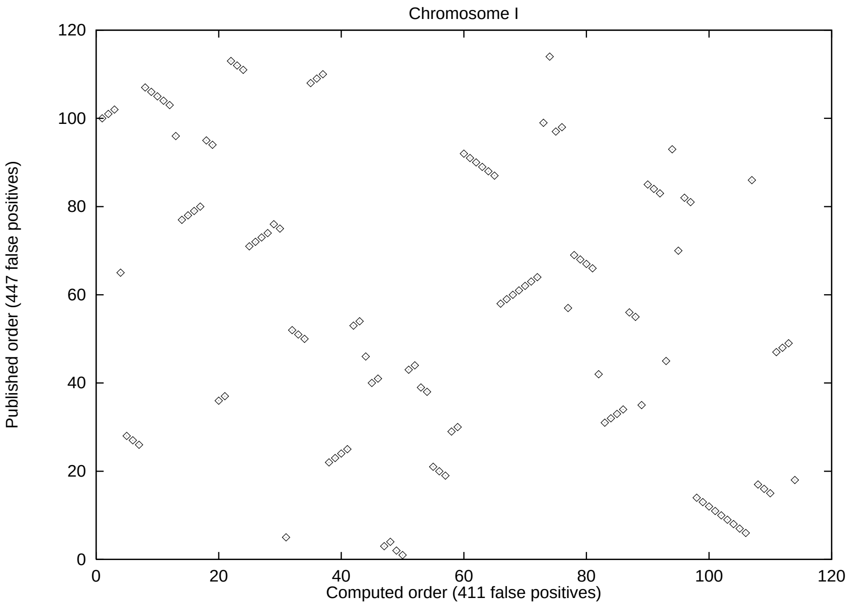

Figure 1: Plot of the published probe order versus our computed probe order for chromosome I of *Aspergillus nidulans*.

Table 3: Average quality of reconstructed maps for simulated data.

	True order		Hamming-TSP			2-COP		
ρ	error	contigs	quality	error	contigs	quality	error	contigs
0.5%	720.1	8.9	0.885	718.4	1.9	0.924	709.7	2.1
0.6%	852.7	8.3	0.865	851.2	1.4	0.895	842.1	2.0
0.7%	998.9	7.5	0.876	997.3	1.7	0.898	988.1	1.8
0.8%	1145.4	7.5	0.877	1144.3	1.5	0.894	1136.1	1.5
0.9%	1272.9	7.1	0.820	1266.5	1.3	0.833	1260.0	1.5
1.0%	1433.4	6.7	0.831	1429.2	1.4	0.842	1420.5	1.5

Table 4: Experiments on simulated data with varying coverage.

		True order		Hamming-TSP			2-COP		
ρ	coverage	error	contigs	quality	error	contigs	quality	error	contigs
0.2%	15	89.6	3.6	0.937	99.5	3.0	0.998	88.1	2.1
0.5%	15	215.9	2.9	0.920	227.4	1.8	0.975	213.4	1.1
0.7%	15	287.8	2.4	0.905	295.6	1.5	0.948	284.1	1.0
0.2%	10	123.2	6.7	0.931	127.4	4.0	0.948	119.6	3.6
0.5%	10	314.1	5.8	0.910	318.6	1.9	0.948	308.3	1.5

maximum-likelihood formulation corresponds to a continuous optimization problem with a nonlinear objective function. As the objective is convex in most regions of interest, it can be evaluated for a fixed probe order by a greedy algorithm such as gradient descent, but finding a good probe order by local search involves many evaluations of this expensive objective. The original motivation of the present work was to study a combinatorial simplification of the objective that could be tackled by discrete techniques, and whose solution would hopefully provide a near-optimal probe order for the original continuous problem. It will be interesting to investigate whether use of our branch-and-cut code yields acceptable probe orders for the full problem of reconstructing orders and spacings.

9 Acknowledgements

We wish to thank Jonathan Arnold for his invaluable assistance, in introducing us to the problem of mapping under the sampling without replacement protocol, and for making his data available on *Aspergillus nidulans*.

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

[5] K. Booth and G. Lueker. "Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms." *Journal of Computer and System Sciences* 13, 335–379, 1976.

[6] Chegireddy, C.R. and H.W. Hamacher. "Algorithms for finding k-best perfect matchings." *Discrete Applied Mathematics* 18, 155–165, 1987.

[7] 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.

[8] Christof, T. *Low-Dimensional 0/1-Polytopes and Branch-and-Cut in Combinatorial Optimization*. Ph.D. thesis, Department of Applied Mathematics, Universität Heidelberg, Germany, Aachen: Shaker, 1997.

[9] 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.

[10] Christof, T., M. Oswald and G. Reinelt. "Consecutive ones and a betweenness problem in computational biology." Proceedings of the 6th Conference on *Integer Programming and Combinatorial Optimization*, Springer-Verlag Lecture Notes in Computer Science 1412, 213–228, 1998.

[11] Duffin, R.J. "On Fourier's analysis of linear inequality systems." *Mathematical Programming* 1, 1974.

[12] 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.

[13] Fasulo, D., T. Jiang, R.M. Karp and N. Sharma. "Constructing maps using the span and inclusion relations." Proceedings of the 2nd ACM *Conference on Computational Molecular Biology*, 64–73, 1998.

[14] Grigoriev, A., H. Lehrach and R. Mott. "An algorithm to detect chimeric clones and random noise in genomic mapping." *Genomics* 22, 482–486, 1994.

[15] 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.

[16] Hamacher, H.W. and M. Queyranne. "K best solutions to combinatorial optimization problems." *Annals of Operations Research* 4, 123–143, 1986.

[17] 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.

[18] 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.

[19] Jünger, M., G. Reinelt and S. Thienel. "Optimal and provably good solutions for the symmetric traveling salesman problem." *Zeitschrift für Operations Research* 40, 183–217, 1994.

[20] Jünger, M. and S. Thienel. "Introduction to ABACUS - A Branch-And-CUt System." *Operations Research Letters* 22, 83-95, 1998.

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

[22] Kececioglu, J., S. Shete and J. Arnold. "Reconstructing order and distance in physical maps using nonoverlapping probes." Technical Report 98-8, Department of Computer Science, The University of Georgia, October 1998.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Naddef, D. and S. Thienel. "Efficient separation routines for the symmetric traveling salesman problem I: general tools and comb separation." Technical report LMC-IMAG, Grenoble, 1999.

[28] Naddef, D. and S. Thienel. "Efficient separation routines for the symmetric traveling salesman problem II: Separating multi handle inequalities." Technical report LMC-IMAG, Grenoble, 1999.

[29] Parida, L. and B. Mishra. "Partitioning k clones: Hardness results and a practical solution to the k-populations problem." Proceedings of the 2nd ACM *Conference on Computational Molecular Biology*, 192–201, 1998.

[30] 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.

[31] 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.

[32] Thienel, S. *ABACUS: A Branch-And-CUt System.* Ph.D. thesis, Department of Applied Mathematics, Universität zu Köln, Germany, 1995.

[33] 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.

[34] 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.

[35] 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 Properties of the 2-consecutive-ones polytope

This appendix uses some common techniques and notations from the area of polyhedral combinatorics. We assume here that the reader is familiar with the basic concepts of that field.

Lemma 1 *The following properties hold.*

(i) $P_{2COP}^{(m,n)}$ *is full-dimensional:*

$$\dim P_{2COP}^{(m,n)} = mn.$$

(ii) *The nonnegativity constraints* $x_{ij} \geq 0$ *define facets for* $P_{2COP}^{(m,n)}$.

(iii) *The trivial constraints* $x_{ij} \leq 1$ *define facets for* $P_{2COP}^{(m,n)}$.

(iv) *For a facet-defining inequality* $f^T x \leq f_0$, *which is not a nonnegativity constraint* $x_{ij} \geq 0$, *it holds that* $f \geq 0$, *and* $f_0 > 0$.

(v) *Let* $m_1 < m_2$ *and* $n_1 < n_2$. *If* $f^T x \leq f_0$ *is valid for* $P_{2COP}^{(m_1,n_1)}$, *then the inequality* $g^T x \leq g_0$ *with*

$$g_{ij} = \begin{cases} f_{ij}, & \text{if } 1 \leq i \leq m_1 \text{ and } 1 \leq j \leq n_2; \\ 0, & \text{otherwise.} \end{cases}$$

and $g_0 = f_0$ *is valid for* $P_{2COP}^{(m_2,n_2)}$. *(Thus trivial lifting holds.)*

(vi) *Let* $m_1 < m_2$ *and* $n \geq 1$ *be given. Let* $f^T x \leq f_0$ *be facet-defining for* $P_{2COP}^{(m_1,n)}$, *and let* $g^T x \leq g_0$ *be an inequality for* $P_{2COP}^{(m_2,n)}$ *obtained by trivial lifting. Then* $g^T x \leq g_0$ *is facet-defining for* $P_{2COP}^{(m_2,n)}$.

(vii) *The inequality* $\sum_{j=1}^n x_{ij} \leq 2$ *is facet-defining for* $P_{2COP}^{(m,n)}$, *for* $1 \leq i \leq m$.

Proof sketch

(i) The mn unit vectors e_{ij} and the 0-vector are in $P_{2COP}^{(m,n)}$ and form an affinely independent set.

(ii) This follows from the fact that the 0-vector and $mn-1$ unit vectors satisfy inequality $x_{ij} \geq 0$ as an equation and are affinely independent.

(iii) A linearly independent set of mn vectors which satisfies the inequality $x_{ij} \leq 1$ with equality is given by the unit vector e_{ij} and the $mn-1$ vectors which are obtained from the unit vector e_{ij} by setting in each vector at a different position one additional entry to 1.

(iv) Since $0 \in P_{2COP}^{(m,n)}$, for all valid inequalities $f^T x \leq f_0$, it follows that $f_0 \geq 0$. Assume that $f^T x \leq f_0$ is not a nonnegativity constraint which defines a facet for $P_{2COP}^{(m,n)}$, and for some k and l, we have $f_{kl} < 0$. In any set of mn affinely independent feasible incidence vectors that satisfy that $f^T x \leq f_0$ as equation, there must be some vector $\overline{x}$ with $\overline{x}_{kl} = 1$. (Otherwise the face induced by $f^T x \leq f_0$ would be contained in the facet $f_{kl} = 0$.) For the vector $\overline{x}'$ that we obtain from $\overline{x}$ by setting $\overline{x}'_{ij} = 0$, we have $f^T \overline{x}' > f_0$. Because $\overline{x}' \in P_{2COP}^{(m,n)}$, this is a contradiction to the basic assumption.

(v) Let $x \in P_{2COP}^{(m_2,n_2)}$ be any 0/1-vector. It is obvious that x restricted to the variables $x_{ij}, 1 \leq m_1, 1 \leq j \leq n_1$, is in $P_{2COP}^{(m_1,n_1)}$. Since $f^T x \leq f_0$ is valid for that restricted vector, and since $g_{ij} = 0$, for $i \geq m_1, j \geq n_1$, the inequality $g^T x \leq g_0$ is valid for $P_{2COP}^{(m_2,n_2)}$.

(vi) Let a set X of $m_1 n$ affinely independent binary vectors such that each vector $x \in X$ is in the facet induced by $f^T x \leq f_0$. A set of $m_2 n$ affinely independent binary vectors which all are feasible for $P_{2COP}^{(m_2,n)}$, and which are all contained in the face induced by the lifted inequality, is obtained as follows. The first $m_1 n$ vectors are the elements of X, filled in the additional components by 0. Then take a vector $x \in X$. The additional vectors are equal to x on the components $ij, i \leq m_1, 1 \leq j \leq n$. For the additional for $m_2 n - m_1 n$ components we set exactly one 1 for each vector in a way that all vectors are distinct.

(vii) Because of (vi), it is sufficient to prove the claim for $m = 1$. For $m = 1$, it is straightforward to see that n linearly independent vectors satisfy the inequality.

□

We can formally define the *k-degree inequalities* as follows. We set $N = \{1, \ldots, n\}$ and $M = \{1, \ldots, m\}$.

Definition 2 Let $m \geq 3, n \geq 4, 3 \leq k \leq m, c \in N$, and let $r_1, \ldots, r_k \in M$ be pairwise disjoint, and let $c_1, \ldots, c_k \in N \setminus \{c\}$ be pairwise disjoint. Then

$$\sum_{i=1}^{k} x_{r_i c} + x_{r_i c_i} \leq k + 2$$

is the k-degree inequality. □

Theorem 1 *The k-degree inequalities are facet-defining for* $P_{2COP}^{(m,n)}$. □

The *augmented k-degree inequalities* are an even larger class that generalize the k-degree inequalities. We give no proof of Theorem 1, as it follows directly from Theorem 2 below.

Definition 3 Let $m \geq 3, n \geq 4, 3 \leq k \leq m, c \in N$ and let $p, q \geq 0, p+q=k$. If $p>0$, then let $r_1, \ldots, r_p \in M$ with $r_i \neq r_j$ for $i \neq j$, and let $c_1, \ldots, c_p \in N \setminus \{c\}$ with $c_i \neq c_j$ for $i \neq j$. If $q>0$, then let $s_1, \ldots, s_q \in M$ with $s_i \neq s_j$ for $i \neq j$, and let $D_1, \ldots, D_q \in N \setminus \{c\}$ with $D_i \cap D_j = \emptyset$, for $i \neq j$. In addition, if $p>0$ and $q>0$, then let $\{c_i\} \cap D_j = \emptyset$ for all $1 \leq i \leq p$ and $1 \leq j \leq q$, and $s_i \neq r_j$, for all $1 \leq i \leq p, 1 \leq j \leq q$. Moreover, we assume that $\{c_1, \ldots, c_p\} \cup \cup_{i \leq q} D_i = N \setminus \{c\}$. Then

$$\sum_{i=1}^{p}(x_{r_i c} + x_{r_i c_i}) + \sum_{i=1}^{q}(2x_{s_i c} + \sum_{j \in D_i} x_{s_i j}) \leq p + 2q + 2$$

is the augmented k-degree inequality. □

Note that for $q=0$, an augmented k-degree inequality is a basic k-degree inequality.

Theorem 2 *The following properties hold.*

(i) *The augmented k-degree inequalities are valid for* $P_{2COP}^{(m,n)}$.

(ii) *The augmented k-degree inequalities are facet-defining for* $P_{2COP}^{(m,n)}$.

Proof sketch

(i) Let $f^T x \leq f_0$ be the augmented k-degree inequality. Consider a binary vector $\overline{x} \in P_{2COP}^{(m,n)}$. It is easy to see that for at most two rows i we can have $\sum_{j \in N} f_{ij} \overline{x}_{ij} = 2$, if $i \in \{r_1, \ldots, r_p\}$, and $\sum_{j \in N} f_{ij} \overline{x}_{ij} = 3$, if $i \in \{s_1, \ldots, s_q\}$. For all other rows we have $\sum_{j \in N} f_{ij} \overline{x}_{ij} \leq 1$, if $i \in \{r_1, \ldots, r_p\}$, and $\sum_{j \in N} f_{ij} \overline{x}_{ij} \leq 2$, if $i \in \{s_1, \ldots, s_q\}$. From this observation, one can conclude $f^T \overline{x} \leq 2q + p + 2$.

(ii) Let $f^T x \leq f_0$ be the augmented k-degree inequality with induced face $F = P_{2COP}^{(m,n)} \cap \{x \mid f^T x = f_0\}$, and let $g^T x \leq g_0, g \neq 0$ be a valid inequality with induced face G such that $F \subseteq G$. For this purpose show that $g = \alpha f$, and $g_0 = \alpha f_0$, for some $\alpha > 0$. We consider several pairs of binary vectors x and y which are all in F and hence all in G. Each pair of vectors differs only slightly in the way we describe. The existence of such vectors is easy to see. The equation $g^T x - g^T y = 0$ implies certain relations on the coefficients of g.

- For $1 \leq r_i$, the first two binary vectors x and y differ by $x_{r_i c} = 1, x_{r_i c_i} = 0$ and $y_{r_i c} = 0, x_{r_i c_i} = 1$. This implies $g_{r_i c} = g_{r_i c_i}$, for all $1 \leq r_i \leq p$.
- Given $i, j \leq p$, x and y differ in $x_{r_i c} = 1, x_{r_j c} = 0$ and $y_{r_i c} = 0, y_{r_j c} = 1$. This implies $g_{r_i c} = g_{r_j c}$, for all $1 \leq i, j, \leq p$.
- Let $1 \leq i \leq q$ and $j_1, j_2 \in D_i$. $x_{s_i j_1} = 1, x_{s_i j_2} = 0$, and $y_{s_i j_1} = 0, y_{s_i j_2} = 1$ implies $g_{s_i j_1} = g_{s_i j_2}$.
- Let $1 \leq i \leq q$. $x_{s_i c} = 1, x_{s_i j_1} = x_{s_i j_2} = 0$, for $j_1, j_2 \in D_i$, and $y_{s_i c} = 0, y_{s_i j_1} = y_{s_i j_2} = 1$ implies $g_{s_i c} = g_{s_i j_1} + g_{s_i j_2}$.
- Let $1 \leq i \leq p, 1 \leq k \leq q, j \in D_k$. $x_{s_k c} = 1, x_{r_i c} = x_{s_k j} = 0$ and $y_{s_k c} = 0, y_{r_i c} = y_{s_k j} = 1$ implies $g_{s_k c} = g_{r_i c} + g_{s_k j}$.

This alltogether implies $g_{ij} = \alpha$, for a positive number α, for all ij for which we have $f_{ij} = 1$, and in addition, we have $g_{ij} = 2\alpha$ for $f_{ij} = 2$. Moreover, we can conclude that $g_0 = \alpha f_0$. Finally we have to show that $g_{ij} = 0$, for the remaining coefficients. Assume that $g_{ij} = \beta > 0$ for a certain ij. Then we can easily find a feasible binary vector $\overline{x}$ with $g^T \overline{x} = g_0 + \beta$, which is a contradiction.

□

The polytope $P_{2COP}^{(m,n)}$ that we considered so far is defined on mn variables. However, in the integer programming formulation we introduced variables x_{ij} only for entries ij for which we have $h_{ij} = 1$, where $H = (h_{ij})$ is the hybridization matrix. Let $S \subseteq M \times N$ be the set of variables corresponding to the non-zero entries of H. The polytope P_{2COP}^{S} associated with the set S (or with a current problem instance) is the orthogonal projection of $P_{2COP}^{(m,n)}$ on the subspace $\{x_{ij} = 0 \mid (i,j) \notin M \times N\}$.

The following lemma states a sufficient condition for a facet-defining inequality of $P_{2COP}^{(m,n)}$ to be facet-defining for P_{2COP}^{S}

Lemma 2 *Let* $f^T x \leq f_0$ *be facet-defining for* $P_{2COP}^{(m,n)}$ *and let* $S \subset M \times N$. *If* $f_{ij} = 0$ *for* $(i,j) \notin M \times N$, *the* $f^T x \leq f_0$ *(restricted to the components* S*) is facet-defining for* P_{2COP}^{S}.

Proof sketch Given a linear irredundant description of $P_{2COP}^{(m,n)}$, a complete description of P_{2COP}^{S} can be obtained by Fourier-Motzkin elimination. Applying that algorithm generates $f^T x \leq f_0$ (restricted to the components S) for describing P_{2COP}^{S}. While Fourier-Motzkin elimination may produce redundant inequalities, a sufficient condition for an obtained inequality to be irredundant [11] holds under the assumptions. The condition is that the original inequality has a zero coefficient for the processed variable. □

In our practical computations we used only augmented k-degree inequalities for which the condition of Lemma 2 holds. Hence, all inequalities which we use are facet-defining for the polytope associated with the given problem instance.