Creating Physical Maps With Automatic Capture of Hybridization Data

David Hall[1,2] rdhall@arches.uga.edu
Suchendra Bhandarkar[1] suchi@cs.uga.edu
Jonathan Arnold[2] arnold@uga.edu

[1] Department of Genetics, University of Georgia, Athens, GA 30602, USA
[2] Department of Computer Science, University of Georgia, Athens, GA 30602, USA

Keywords: physical map, hybridization, microcannonical annealing

1 Introduction

Physical maps are a partial ordering of distinguishable chromosomal features. A rapid and relatively inexpensive method for creating them involves fingerprinting clones from a chromosomal library for the presence of probe sequences using hybridization. Hybridization experiments are typically carried out in parallel by fixing DNA's from multiple clones to a filter and incubating that filter with radioactively or fluorescently labeled probe DNA. Clones hybridizing to the probe are identified by spots on the filter where a significant amount of probe DNA remains after the filter is washed.

Originally, the analysis of probed filters was done manually. In modern high-throughput environments it is desirable that capture of hybridization data be automated. To do this, the amount of probe sequence (i.e. radioactivity or fluorescence) detected on different locations of a filter is digitally encoded and software determines which clones hybridize to the probe. Locations on the membrane with levels of radioactivity or fluorescence greater than a threshold value are scored as positive with regard to hybridization. Locations with levels less than the threshold are scored as negative.

In this study we examine several methods for creating physical maps with the automatic capture of data. In one method hybridization data is scored by software using a threshold. Physical maps are then generated using two algorithms, the Hamming distance traveling salesman algorithm (HDTSP) [3] and the maximum likelihood algorithm of Alizadeh *et al.* (ML) [1]. A novel approach for creating physical maps using unscored hybridization data is also evaluated. In this approach *raw* levels of hybridization are used by a novel physical mapping algorithm that will be described. This algorithm can be used for creating physical contig maps when probes are unique and clones and probes are of a uniform size.

2 Method and Results

An algorithm for ordering unique probes using unscored hybridization data is first presented. This algorithm can use hybridization data consiting of real numbers. It is assumed that larger numbers denote stronger hybridization. If data is noise-free, then hybridization values corresponding to overlapping clones and probes should be large and values corresponding to non-overlapping clones and probes should be small. Let m denote the number of clones and n denote the number of probes. Let A be a $m \times n$ matrix where $a_{i,j}$ gives the hybridization value for the i^{th} clone and the j^{th} probe. If clones and probes are of uniform size and each probe samples a unique chromosomal interval, then each clone will overlap with at most two probes. If data is noise-free and the columns of A are permuted so that they correspond to the correct order of probes, then in each row the two largest values will be in adjacent positions. If this matrix is converted to binary by scoring with a perfect threshold, the resulting binary matrix will have the consecutive ones property [2].

An objective function based on the noise free matrix described above follows:

$$\sum_{i=1}^{m} \max_{j} \{a_{i,j} + a_{i,j+1}\} \tag{1}$$

If data is noise free and the columns of A are permuted so they correspond to the correct probe order, then this function will be maximum. In this study the microcannonical annealing algorithm was used to search for a permutation of columns maximizing the function. We will denote the resulting algorithm MCA-1.

The approaches using scored data mentioned in the introduction and this novel approach using unscored data were evaluated by simulation with data sampled from the *Aspergillus nidulans* physical mapping project. These data consist of grayscale raster image files created from radioactively probed filters. Average grayscale values were determined for each clone in the negative of these images. The values for each filter were then corrected for variation between images by normalizing the values to the range (0,1), where 0 was equal to the value of the clone with the smallest greyscale level on the image, and 1 was equal to the greatest. 2 megabase chromosomes with different coverages of 40 kilobase clones were generated *in silico* by a Poison process. A maximum set of non-overlapping probes was selected from among the clones and designated probes. Hybridization values for clones and probes that overlap *in silico* were selected from hybridization values of *A. nidulans* clones and probes known to overlap. Similarly, hybridization values were generated for non-overlapping simulated clones and probes. For the HDTSP and ML algorithms, hybridization data were scored using several different thresholds.

Probes were ordered using the HDTSP, ML, and the MCA-1 algorithms. Results were evaluated by determining the percentage of adjacencies common to both the true order of probes and the computed ordering (adjacency quality). Each entry in the following table shows the mean of 100 independent data sets.

Adjacency Quality (%)								
	coverage = 5				coverage = 15			
algorithm	threshold				threshold			
	0.01	0.05	0.15	0.25	0.01	0.05	0.15	0.25
HDTPS	32	58	75	67	36	88	84	72
ML	17	55	75	66	70	91	90	88
MCA-1	82				100			

It is apparent from this data that the choice of a threshold value is important. Also the threshold value giving the best result for the HDTPS and ML algorithms is dependent on the coverage. This suggests that in practice it may be difficult to choose the best threshold value. The MCA-1 algorithm outperformed the other two at all threshold values examined, suggesting this approach may be preferable in the context of automatic data capture.

References

[1] Alizadeh, F., Karp, R.M., Weisser, D.K. and Zweig, G., Physical mapping of chromosomes using unique probes, *J. Comput. Biol.*, 2(2):159–184, 1995.

[2] Booth, K.S. and Leuker, G.S., Testing for consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms, *J. Comp. and Syst. Sci.*, 13:335–379, 1976.

[3] Cuticchia, A.J., Arnold, J. and Timberlake, W.E., The use of simulated annealing in chromosome reconstruction experiments based on binary scoring, *Genetics*, 132(2):591–601, 1992.