Aligning Optical Maps

Yu-Chi Liu [1][2], Michael S. Waterman [1], Anton Valouev [1], Lei Li [1], Yu Zhang [1], Yi Yang [1], Jong-Hyun Kim [1], David C. Schwartz[3]

Keywords: optical mapping, restriction maps alignment, dynamic programming

1 Introduction.

Optical Mapping is a single molecule system for the construction of ordered DNA restriction maps [4]. It uses light microscopy to directly image individual DNA molecules bound to charged surfaces, cleaved by restriction endonucleases. Cut sites are flagged by small, visible gaps. Notably, restriction fragments retain their original order, and the resulting string of restriction fragment masses is termed an optical map. Shotgun Optical Mapping is an approach that constructs whole-genome DNA restriction maps by overlapping optical maps derived from randomly sheared genomic DNA molecules. Each molecule is mapped and aligned to form contigs consisting of $10 - 50$x coverage of a given locus; as such, these maps have served as scaffolds for sequence assembly and validation [3].

In the map assembly problem, sensitive detection of overlaps between optical maps plays a crucial role, and experimental errors make the analysis of this problem complex. By modifying the dynamic programming algorithm in [2] for restriction maps alignment with suitable scoring function and parameters, we are able to find accurate matchings between two overlapping optical maps under the limits of known experimental errors.

2 Data.

Due to experimental condition, the measured maps have the following features:

I. *missing cuts*,

which occur when molecules are not cleaved at all restriction sites.

In one data set, the probability of missing cut at a cut site is about 0.2.

II. *false cuts*,

which occur when there are cleavages detected which are not at a restriction site.

We have about 5 false cuts per Mb in one data set.

III. *sizing errors*,

which are imprecisions in measuring the length of restriction fragments.

Suppose l_r is the true fragment length and l_m is what we observed in the data, then $|l_m - l_r|$ is modeled to be proportional to $\sqrt{l_r}$.

3 Methods.

Let a *segment* $[i, k]$ of a map consists sites i through k. Let a *matching pair* between two maps be denoted by $(i, j; k, l)$.

[1] Program of Molecular and Computational Biology, University of Southern California, Los Angeles, California, 90089.

[2] E-mail: `ycliu@usc.edu`

[3] Laboratory for Molecular and Computational Genomics, Department of Chemistry and Laboratory of Genetics, University of Wisconsin-Madison, Madison, Wisconsin 53706. E-mail: `dcschwartz@facstaff.wisc.edu`

Suppose we have a global alignment Π between a map A with m sites and the other map B with n sites is a sequence of ordered matching pairs $(i_1, j_1; k_1, l_1)$ $(i_2, j_2; k_2, l_2)$ $\ldots$ $(i_d, j_d; k_d, l_d)$, where $k_t < i_{t+1}$ and $l_t < j_{t+1}$ for each $t < d$. Let q_x denote the position of site x on map A, and r_y the position of site y on map B. With $\lambda \geq 0$ and $\nu \geq 0$, the score of Π is defined as:

$$\begin{aligned} \mathit{score}(\Pi) = \sum_{t=1}^{d} \sigma(i_t, j_t; k_t, l_t) + l(q_{i_1}, r_{i_1}) + \sum_{t=2}^{d} l((q_{i_t} - q_{k_{t-1}}), (r_{j_t} - r_{l_{t-1}})) \\ + l((q_m - q_{k_d}), (r_n - r_{l_d})) - \lambda \left[m + n - \sum_{t=1}^{d} (k_t - i_t + 1) - \sum_{t=1}^{d} (l_t - j_t + 1) \right], \end{aligned}$$

where

$$\begin{aligned} \sigma(i_t, j_t; k_t, l_t) = \nu \cdot (\#\,\text{of matching sites pair in segment}\,[i_t, j_t; k_t, l_t]) \\ + l((q_{k_t} - q_{i_t}), (r_{l_t} - r_{j_t})) - \lambda((k_t - i_t) + (l_t - j_t)). \end{aligned}$$

Each pair of matching sites in segment is rewarded by ν in $\sigma(i_t, j_t; k_t, l_t)$. It also takes care of random permutation of restriction sites position. Each site not matched with any others (due to a false cut or a missing cut) is penalized by λ. $l(a, b)$ is the scoring function of length similarity for two segments, one with length a and the other with length b. So the function $l(a, b)$ takes care of the distance discrepancy between a matching pair.

To obtain an appropriate function for $l(a, b)$, we first, using a central limit theorem, model the measurement distribution (conditional on the true fragment length). Then we derive a log likelihood function to use as $l(a, b)$.

This approach, which is for global map alignment, can be modified to yield local alignments and overlap alignment. To obtain an algorithm for overlap alignment, we modify the recursion in [2] as follows.

Initialize both the first row and column as 0 when calculating the maximum score matrix of alignments (which means both the open and end gaps are not penalized [1]). We are able to find the best overlap alignment (with highest score) between two optical maps under the scoring scheme we mentioned above.

References

[1] Huang, X. 1992. A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, 14:18-25

[2] Huang, X. and Waterman, M.S. 1992. Dynamic programming algorithms for restriction map comparison. *Comp. Appl. Bio. Sci.*, 8:511-520.

[3] Lim, A., Dimalanta, E.T., Potamousis, K.D., Yen, G., Apodoca, J., Tao, C., Lin, J., Qi, R., Skiadas, J., Ramanathan, A., Perna, N.T., Plunkett, G. 3rd., Burland, V., Mau, B., Hackett, J., Blattner, F.R., Anantharaman, T.S., Mishra, B., Schwartz, D.C. 2001. Shotgun optical maps of the whole Escherichia coli O157:H7 genome. *Genome Res.*, 11(9):1584-93.

[4] Schwartz, D.C., Li, X., Hernandez, L.I., Ramnarain, S.P., Huff, E.J., Wang, Y.K. 1993. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, 262:110-114.