Statistical Algorithms and Software for Genomics *

Thomas Anantharaman and Bud Mishra

Computer Science Department, Courant Institute,
New York University, New York, NY 10012-1185.

August, 1997

1 Introduction

There are many large system problems that are hard to model exactly or in a computationally tractable fashion. Examples include the mapping of human DNA, speech recognition, and automated learning in computer chess. Traditional artificial intelligence solution techniques for such problems rely on a combination of custom encoding of expert knowledge and heuristic search. They take much time to hand craft and then often are unable to take advantage of faster computers as they become available. In this context, we explore the advantage of using *Statistical Search* techniques in which the knowledge is encoded in some form of statistical model whose parameters are automatically adjusted or trained with domain data. The benefits are faster development times, greater solution accuracy (compared to hand crafted solutions) and the ability to allow the problem size and desired solution accuracy to be scaled up with computational resources.

We apply this approach to certain critical computational problems in mapping the human genome. We use a *Bayesian model* to provide the best solution accuracy as a function of the number of parameters. Heuristic search techniques derived from artificial intelligence are used to search the model space in an efficient manner in the average case.

2 Preliminary

The computational problem to be solved is a key step in producing a *physical map of the human genome* using the *optical mapping* technique (See [SLH93]). The human genome consists of 46 DNA strands called chromosomes. The ultimate object of the Human Genome project is to decode the entire DNA base-pair (abbreviated bp) sequence. This sequence totals about 3 billions bp's and current techniques make decoding the entire DNA sequence in a single step unrealistic in the near future. Instead a preliminary goal is to produce a physical map of landmarks, which is a sampled version of the entire base-pair sequence.

The landmarks used by *optical mapping* are short 6 bp sequences of DNA called restriction sites that are digested (broken) by restriction enzymes and occurring about every 4000 bp's along the DNA. The starting point for constructing a physical map of the human DNA is typically a collection of random human DNA pieces (called a clone library). Each DNA sequence in the clone library is reproducible using genetically engineered organisms into which the human DNA sequence is inserted and reproduced (cloned). The most popular clone library at this time is based on the Yeast Artificial Chromosomes (YACs) or the related Bacterial Artificial Chromosomes (BACs), and consists of DNA sequences of about 150 kbp (kilo base pairs) length on average. The original location on the human genome of these sequences is unknown, but the library is grown to contain a sufficient number of them to cover the entire human genome about 5 times (5X coverage). The idea is to produce a physical map for each DNA sequence (or molecule) in the clone library, and then combine them into larger physical maps by inference from the common overlap regions.

One very promising technique to build physical maps of DNA molecules from the BAC based clone library is *optical mapping* [SLH93, S95]. Optical mapping uses fluorescent microscopy to observe the DNA molecules under an optical microscope. The DNA molecules adhere to a glass surface and are treated with restriction enzymes and a suitable fluorescent dye. The locations at which the DNA molecules are digested (the restriction sites referred to earlier) by a particular restriction enzyme are visible as gaps and can be used to compute a restriction map of the DNA molecules. One advantage of optical mapping is that it can be automated and therefore carried out rapidly.

*This paper was partly supported by an

Proceedings of the COMPSAC '97 - 21st International Computer Software and Applications Conference
0730-3157/97 $10.00 © 1997 IEEE

The key optical mapping computation problem is computing the restriction map from the optical image data described previously. The *computational problem is hard* because of a number of error sources that are unavoidable. We need to consider following categories of errors in image data: 1) Misidentification of spurious materials in the image as DNA, 2) Identifying multiple DNA molecules as one, 3) Identifying partial DNA molecules as complete, 4) Errors in estimating sizes of DNA fragments, 5) Incomplete digestion of DNA, 6) Cuts visible at locations other than digest sites, and 7) Orientation of DNA molecule is not always known.

The published BAC maps [CHYS95], produced using optical mapping are generated by manually selecting the best molecules in the image, extracting the partial maps of each such molecule with simple image processing software that integrates the illumination of each fragment as a relative estimate of its size in base-pairs, and combining them manually like a jig-saw puzzle. The final map is then obtained by averaging the location of the restriction sites (or the sizes of fragments between restriction sites).

It has provensurprisingly hard to duplicate this map computation in a fully automatic manner. Given the ability of human mappers to solve this problem, albeit at a rate of 1-2 weeks per map of one DNA molecule, this problem can be considered a suitable candidate for AI solution techniques.

3 Bayesian Search Technique

The Bayesian model and a gradient optimization algorithm that havebeen used successfully is summarized here [AMS96], and is similar to algorithms used in maximum likelihood estimation. The following notation will be used to describe the parameters of the independent processes responsible for the statistical structure of the data:

• p_{c_i} = Probability that the i^{th} sequence specific restriction site in the molecule will be visible as a cut. σ_i is the standard deviation of the position of the observed cut when present.

• λ_f = Expected number of false-cuts per molecule observed.

• p_b = Probability that the data is invalid ("bad"). In this case, the data item is assumed to be an unrelated piece of DNA or a partial molecule. The cut-sites (all false) on this data item are assumed to have a Poisson distribution with parameter = λ_n.

By Bayes' rule

$$\mathbb{P}\mathrm{r}[\mathcal{H}|\mathcal{D}] = \frac{\mathbb{P}\mathrm{r}[\mathcal{D}|\mathcal{H}]\ \ \mathbb{P}\mathrm{r}(\mathcal{H})}{\mathbb{P}\mathrm{r}[\mathcal{D}]}$$

Assuming that the prior $\mathbb{P}\mathrm{r}[\mathcal{H}]$ distribution is given in terms of just the number of restriction sites, we wish to find the "most plausible" hypothesis $\mathcal{H}$ by maximizing $\mathbb{P}\mathrm{r}[\mathcal{D}|\mathcal{H}]$.

In our case, $\mathcal{H}$ is simply the final map (a sequence of restriction sites, $h_1, h_2, \ldots, h_N$) augmented by the auxiliary parameters such as p_{c_i}, σ_i, λ_f, etc. When we compare a data item D_j with respect to this hypothesis, we need to consider every possible way that D_j could have been generated by $\mathcal{H}$. In particular we need to consider every possible alignment, where the k^{th} alignment, A_{jk}, corresponds to a choice of the orientation for D_j as well as identifying a cut on D_j with a true restriction site on $\mathcal{H}$ or labeling the cut as a false cut.

As a consequence of the pairwise independence and the preceding discussion, we have the following:

$$\begin{aligned}\mathbb{P}\mathrm{r}[\mathcal{D}|\mathcal{H}] &= \prod_j^M \mathbb{P}\mathrm{r}[D_j|\mathcal{H}],\\ \mathbb{P}\mathrm{r}[D_j|\mathcal{H}] &= ((1-p_b)/2)\sum_k \mathbb{P}\mathrm{r}_{jk} + p_b e^{-\lambda_n}{\lambda_n}^{M_j}\end{aligned}$$

where j and k range overthe data set and the alignments, respectively, and $\mathbb{P}\mathrm{r}_{jk} = \mathbb{P}\mathrm{r}[D_j^{(k)}|\mathcal{H}]$, with $D_j^{(k)}$ denoting the "interpretation of D_j with respect to the alignment A_{jk}." In the most general case, we proceed as follows. Let

$N \equiv$ Number of cuts in the hypothesis $\mathcal{H}$.

$h_i \equiv$ The ith cut location on $\mathcal{H}$.

$M_j \equiv$ Number of cuts in the data D_j.

$K_j \equiv$ Number of possible alignments of the data D_j against the hypothesis $\mathcal{H}$ (or its reversal, the flipped alignment $\mathcal{H}^R$).

$s_{ijk} \equiv$ The cut location in D_j matching the cut h_i in $\mathcal{H}$, given the alignment A_{jk}.

$m_{ijk} \equiv$ An indicator variable, taking the value 1 iff the cut s_{ijk} in D_j matches a cut h_i in the hypothesis $\mathcal{H}$, given the alignment A_{jk}. It takes the value 0, otherwise.

$F_{jk} \equiv$ Number of false (non-matching) cuts in the data D_j for alignment A_{jk}, that do not match any cut in the hypothesis $\mathcal{H}$.

Using this notation one can express $\mathbb{P}\mathrm{r}_{jk}$ as follows:

$$\begin{aligned}\mathbb{P}\mathrm{r}_{jk} &= \prod_{i=1}^N \left[p_{c_i}\frac{e^{-(s_{ijk}-h_i)^2/2\sigma_i^2}}{\sqrt{2\pi}\sigma_i}\right]^{m_{ijk}}\\ &\quad\times \prod_{i=1}^N [1-p_{c_i}]^{1-m_{ijk}}\ \ e^{-\lambda_f}\ {\lambda_f}^{F_{jk}} \qquad (1)\end{aligned}$$

Proceedings of the COMPSAC '97 - 21st International Computer Software and Applications Conference
0730-3157/97 $10.00 © 1997 IEEE

The log-likelihood can then be computed as: $\mathcal{L} \equiv \sum_j \log \mathbb{P}\mathrm{r}[D_j|\mathcal{H}]$. The *prior* $\mathbb{P}\mathrm{r}[\mathcal{H}]$ used is a function of N only and takes the shape of a Poisson distribution parameterized by the expected number of restriction sites, which in theory can be estimated by a knowledge of the restriction site size and DNA molecule size as described in [W95], but in practice the average number of actual observed restriction sites in the data is a better estimate.

3.1 Gradient Search Algorithm

In order to find the most plausible restriction map, we shall optimize the cost function derived earlier, with respect to the following parameters: cut sites $= h_1, h_2, \ldots, h_N$, cut rates $= p_{c1}, p_{c2}, \ldots, p_{cN}$, standard deviation of cut sites $= \sigma_1, \sigma_2, \ldots, \sigma_N$, and auxiliary parameters $= p_b, \lambda_f$ and λ_n.

Here we just list the set of update equations that can be iterated to find such an extremal point [AMS96]. We first define the following:

$$\pi_{jk} \equiv ((1-p_b)/2)\ (\mathbb{P}\mathrm{r}_{jk}/\mathbb{P}\mathrm{r}_j)$$

This is the relative probability of the alignment A_{jk} for data item D_j. We write $\Psi_{\alpha i}$ to denote

$$\begin{array}{rlll} & \Psi_{\alpha i} & \equiv & \sum_j \sum_k \pi_{jk} m_{ijk} s^{\alpha}_{ijk}, \quad \alpha = 0, 1, \text{ or } 2. \\ \text{Also,} & \mu_g & \equiv & \sum_j \sum_k \pi_{jk} \\ & \gamma_g & \equiv & \sum_j \sum_k \pi_{jk} M_j \end{array}$$

Using this notation the update rules are:

$$h_i := (\Psi_{1i}/\Psi_{0i}). \tag{2}$$

For $p_c = p_{c1} = \cdots = p_{cN}$ (*Constrained*)

$$p_c := ((\sum_i \Psi_{0i}/N)/\mu_g). \tag{3}$$

This constraint is easily justified for physical reasons. Similarly for $\sigma = \sigma_1 = \cdots = \sigma_N$ (*Constrained*)

$$\sigma^2 := ((\sum_i \Psi_{2i} - \Psi^2_{1i}/\Psi_{0i})/\sum_i \Psi_{0i}). \tag{4}$$

$$\lambda_f := (\gamma_g/\mu_g) - \sum_i (\Psi_{0i}/\mu_g). \tag{5}$$

(Note that the molecules are already normalized to unit length.)

3.2 Update Algorithm: (DP)

In each update step, we need to compute the new values of the parameters based on the old values of the parameters, which affect the "moment functions:" Ψ_{0i}, Ψ_{1i}, Ψ_{2i}, μ_g and γ_g. For the ease of expressing the computation, we shall use additional auxiliary expressions as follows:

$$\begin{aligned} \mathtt{P}_j &\equiv \sum_k (\mathbb{P}\mathrm{r}_{jk}/e^{-\lambda_f}), \\ \Phi_{\alpha ij} &\equiv \sum_k (\mathbb{P}\mathrm{r}_{jk} m_{ijk} s^{\alpha}_{ijk}/e^{-\lambda_f}), \end{aligned}$$

where $\alpha = 0$, 1, or 2. Note that, the original moment functions can now be computed in terms of these auxiliary parameters. For instance:

$$\begin{aligned} \mathbb{P}\mathrm{r}_j &= ((1-p_b)/2)e^{-\lambda_f} \times \mathtt{P}_j + p_b e^{-\lambda_n} {\lambda_n}^{M_j} \\ \Psi_{\alpha i} &= ((1-p_b)/2)e^{-\lambda_f} \sum_j (\Phi_{\alpha ij}/\mathbb{P}\mathrm{r}_j) \end{aligned}$$

($\alpha = 0$, 1, or 2). The definitions for $\mathtt{P}_j$ and $\Phi_{\alpha ij}$ involve *all alignments* between each data element D_j and the hypothesis $\mathcal{H}$. This number is easily seen to be exponential in the number of cuts N in the hypothesis $\mathcal{H}$, *even if one excludes such physically impossible alignments as the ones involving cross-overs* (i.e., alignments in which the order of cuts in $\mathcal{H}$ and D_j are different). First we present a recurrence equation for computing $\mathbb{P}\mathrm{r}_j$:

$$\begin{aligned} \mathtt{P}_{q,r} \equiv\ & \lambda_f \mathtt{P}_{q+1,r} + \sum_{t=r}^{N} \mathtt{P}_{q+1,t+1} \left\{ \prod_{i=r}^{t-1} (1-p_{ci}) \right\} \\ & \times p_{ct}\ \frac{e^{-(h_t - s_q)^2/2\sigma_t^2}}{\sqrt{2\pi}\sigma_t}, \end{aligned} \tag{6}$$

where $1 \le q \le M_j$ and $1 \le r \le N+1$. The recurrence terminates in $\mathtt{P}_{M_j+1,r} = \prod_{i=r}^{n}(1-p_{ci})$ and $\mathbb{P}\mathrm{r}_j = \mathtt{P}_{1,1}$. By computing $\mathtt{P}_{q,r}$ in descending order of r, *only two new terms* needs to be computed for each $\mathtt{P}_{q,r}$, giving a time complexity of $O(M_j N)$. The complexity can be further improved by taking advantage of the fact that the exponential term is negligibly small unless h_t and s_q are sufficiently close (e.g., $|h_t - s_q| \le 3\sigma_t$). It can be shown that this reduces the complexity to $O(\sigma M_j N)$ (where σ is an upper bound on all the σ_t values). Summing over all molecules D_j the total time complexity is $O(\sigma M N)$, where $M = \sum_j M_j$. The space complexity is trivially bounded by $O(M_{max} N)$ where $M_{max} = \max_j M_j$. The recurrence equations for other terms are similar.

3.3 Global Search Algorithm

Recall that our prior distribution $\mathbb{P}\mathrm{r}[\mathcal{D}|\mathcal{H}]$ is multimodal and the local search based on the gradients by itself cannot evaluate the best value of the parameters. Instead, we must rely on some sampling method to find points in the parameter space that are likely

Proceedings of the COMPSAC '97 - 21st International Computer Software and Applications Conference
0730-3157/97 $10.00 © 1997 IEEE

to be near the global maxima. Furthermore, examining the parameter space, we notice that the parameters corresponding to the number and locations of restriction sites present the largest amount of multi-modal variability and hence the sampling may be restricted to just $\bar{h} = (N;\ h_1,\ h_2,\ \ldots,\ h_N)$. The conditional observation probability $\Pr[\mathcal{D}|\mathcal{H}]$ can be evaluated pointwise in time $O(\sigma MN)$ and the nearest local maxima located in time $O(\sigma MN^2)$, though there is no efficient way to sample all local maxima exhaustively.

Thus, our global search algorithm will proceed as follows: we shall first generate a set of samples ($\bar{h}_1$, $\bar{h}_2$, $\bar{h}_3$, ...); these points are then used to begin a gradient search for the nearest maxima and provide hypotheses ($\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$, ...); the hypotheses are then ranked in terms of their posterior probability $\Pr[\mathcal{H}|\mathcal{D}]$ (whose relative values also provide the confidence value of each hypothesis) and the one (or more) leading to maximal posterior probability is presented as the final answer.

We use a global search over an approximate posterior distribution. The approximate score for a map with N cuts is the score for the worst submap with 2 cuts. Scores for all 2 cut maps can be precomputed if the space is discretized (for instance, each $h_i \in \{0,\ 1/D,\ 2/D\ \ldots,\ 1\}$). Branch and bound can now be used to search over the approximate score for larger maps, using the following bounds: (1) If a 2 cut map scores worse than the best N cut map so far, no N cut map with those 2 cuts need be considered; and (2) If the best 2 cut map involving cut h scores worse than the best N cut map so far, no N cut map with cut h need be considered. These two bounds are sufficient to allow exhaustive searches up to about 8 cuts on a Sun-SS20, and larger maps can be searched by fixing some of the best cuts found so far so that no more than 8 cuts are varied at a time.

4 Experimental Results

Pure Lambda DNA (I): Deposited manually using the peel technique. (June 1995).

R. Enz.	Cuts	Conf.	Map Err.
Sca I	6	100%	0.67%
*Ava*I	9	99%	0.83%

Pure Lambda DNA (II): Deposited by a robot as a grid of spots. (July 1996).

R. Enz.	Cuts	Conf.	Map Err.
BamH I	5	37%	1.07%
BamH I	5	100%	0.79%
BamH I	5	100%	0.22%
BamH I	7	9%	Wrong Map

Human DNA Clones: Using a Cosmid Vector, and deposited as a grid of spots. Map verified by contig and Gel as having 6 cuts, with one small fragment ($< 1kB$, and optically undetectable in most of the images) missing [Marked (*) in the table below]. (October 1996).

R. Enz.	Cuts	Conf.	Map Err.
Mlu I	6	50%	(*)
Mlu I	6	100%	(*)
Mlu I	5	72%	1.45%
Mlu I	10	8%	Wrong Map

5 Conclusion

In this paper, we have made three critical contributions towards the solution of the assembly problem for optical mapping data: (1) We provide the first detailed model of the data produced by the optical mapping process and use it in a Bayesian approach that hinges on the fact that the model is complete. (2) We formulate an efficient statistical algorithm that implements the update rules for the model parameters iteratively using dynamic programming. As the multi-modal structure of the prior excludes purely local search algorithms, we necessarily rely on good heuristics (e.g., branch-and-bound). This is consistent with our prior NP-completeness results for this problem [AMS96]. (3) Finally, experiments based on an implementation (in C, running on Sparc 20's) produce highly accurate maps over wide range of experimental variations.

References

[AMS+97] T.S. Anantharaman et al. **Statistical Algorithms for Optical Mapping of the Human Genome**. *1997 Genome Mapping and Sequencing Conference*, Cold Spring Harbor, NY, May, 1997.

[AMS96] T.S. Anantharaman, B. Mishra and D.C. Schwartz. **Genomics via Optical Mapping II: Ordered Restriction Maps**. To appear in *Journal of Computational Biology* 1997.

[CHYS95] W. Cai et al. **Ordered Restriction Endonuclease Maps of Yeast Artificial Chromosomes Created by Optical Mapping on Surfaces**, *Proc. Natl. Acad. Sci., USA*, **92**:5164–5168, 1995..

[S95] A. Samad et al. **Mapping the Genome One Molecule at a Time—Optical Mapping**. *Nature*, **378**:516–517, 1995.

[SLH93] D.C. Schwartz et al. **Ordered Restriction Maps of Saccharomyces Cerevisiae Chromosomes Constructed by Optical Mapping**. *Science*, **262**(5130):110-114, 1993

[W95] M.S. Waterman. **Introduction to Computational Biology**. 1995.

[0] **Acknowledgments** Our thanks go to all the people involved in optical mapping, especially our colleague David Schwartz, the research scientists who provided the data: Virginia Clarke, Junping Jing and Joane Edington and the researchers who provided the underlying systems support: Brett Porter, Ed Huff, Estarose Wolfson, Alex Shenker and Ernest Lee.

Proceedings of the COMPSAC '97 - 21st International Computer Software and Applications Conference
0730-3157/97 $10.00 © 1997 IEEE