Estimating Recombination Rate Distribution by Optimal Quantization

Mingzhou Song
Department of Computer Science
Queens College, Flushing, NY 11367
msong@cs.qc.edu

Stephane Boissinot
Department of Biology
Queens College, Flushing, NY 11367
Stephane_Boissinot@qc.edu

Robert M. Haralick
Doctoral Program in Computer Science
Graduate Center, City University of New York
New York, NY 10016
haralick@gc.cuny.edu

Ihsin T. Phillips
Department of Computer Science
Queens College
Flushing, NY 11367
yun@cs.qc.edu

Abstract

We obtain recombination rate distribution functions for all human chromosomes using an optimal quantization method. This non-parametric method allows us to control over-/under-fitting. The piece-wise constant recombination rate distribution functions are convenient to store and retrieve. Our experimental results showed more abrupt distribution functions than two recently published results. In the previous results, the over-/under-fitting issues were not addressed explicitly. Our estimation had greater log likelihood over a previous result using Parzen window. It suggests that the optimal quantization technique might be of great advantage for estimation of other genomic feature distributions.

1. Introduction

Recombination is the primary biological event pushing evolution. Biologists are interested in a high resolution recombination map that presents accurately how often a recombination event occurs at a specific location in a chromosome. Linear correlations have been established between recombination rate with factors such as Poly(A)/Poly(T) density, CpG density, GC content density, RefSeq gen count, PPY/PPU density, UniGene cluster count [2]; LINE density, SINE density, $(AC)_n$ density and chromosome position [4]. To understand accurately how recombination occurs, a recombination rate distribution (RRD) function can be used to quantify the recombination events at any location in a chromosome. The RRD functions of human chromosomes are first published in the Marshfield map [4]. The Iceland map [2] also obtained RRD functions but from a much larger sample size.

We find a RRD function that best reveals the information contained in the data. We propose to use optimal quantization [3], a non-parametric methodology, to estimate a piecewise constant probability density function (p.d.f.). The estimated function is optimal in that over-/under-fitting are minimized by selecting the best control parameters. Comparisons made with other methods show the advantage of our approach in terms of the cross-validated log likelihood. In addition, the quantized p.d.f. representation is more convenient to use than other kernel methods such as Parzen window or k nearest neighbor, because it can be accessed as a table in logarithm time.

The paper is organized into five sections. In Section 1, we introduce the RRD estimation problem and our strategy. In Section 2, we explain the recombination and review methods to examine recombination events and estimate a RRD function. In Section 3, we describe the theoretical and algorithmic aspect of optimal quantization. In Section 4, we demonstrate RRDs obtained by optimal quantization and compare its performance quantitatively with the Parzen window results. Finally in Section 5, we conclude our study and describe some future work related to biological feature distribution estimation that can be done using optimal quantization.

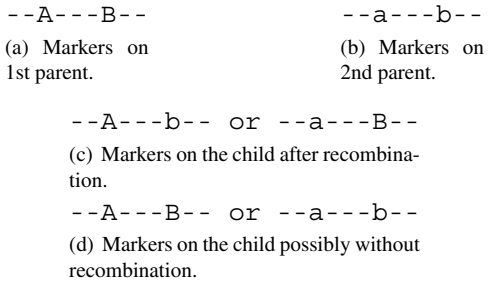
2. Recombination Rate Distribution Function

Recombination plays a central role in molecular revolution. The mechanism of recombination can reveal directly

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

how human evolution might happen. In the nucleus of each human cell except the gamete, there are 46, or 23 pairs of, chromosomes: 22 pairs of autosomes and 1 pair of sex chromosomes [1]. The autosomes are from chromosome 1 to 22. There are two copies of each autosome called *homologous chromosomes*. The two sex chromosomes are either XX or XY. During the reproduction of the gamete, a process called *meiosis*, the chromosomes of the child are obtained by combining half of the chromosomes from one parent with half of the chromosomes from the other parent, that is, combining 22 autosomes plus one sex chromosome from one parent with those from the other parent. Only homologous chromosomes will be combined; the two sex chromosomes always combine themselves. When each pair is combined, the contents of the chromosomes are exchanged at some points along the chromosomes, which could be due to cross-over or gene conversion. Thus the child chromosomes do not necessarily contain exact copies of parent chromosomes. This information exchange between parent chromosomes is called *recombination*. *Recombination rate* is defined as the number of recombination events in a unit length of chromosome in terms of base pairs, usually in centiMorgan per Mbps (cM/Mb). The RRD function maps a location on the chromosome to a recombination rate value. However, experimental data on recombination are still very limited due to the cost and complexity of experiments. As only recently the complete human genome physical map becomes available, an accurate quantitative representation of the RRD becomes possible and is under heavy investigation.

In practice, recombination events are identified using both genetic and physical maps. On a genetic map, there are markers with each one being a unique feature. A marker has two or multiple forms, called *alleles*. The alleles can be identified quickly by polymerase chain reaction (PCR). Locations of markers on the physical map are determined in advance. With markers and their locations on the physical map, a recombination event might be identified in a practical manner without sequencing the whole genome. The resolution of the identified events increases with the number of markers used. This method is illustrated in Fig. 1. The first parent has markers as shown in Fig. 1(a) and the second parent has the same markers but with different alleles in Fig. 1(b). If the child has the markers shown in Fig. 1(c), then at least one recombination event has occurred at some location between the markers A and B. If the child has the same alleles as their parents in Fig. 1(d), then it is unlikely to have a recombination event between A and B if the markers are close enough. This method cannot detect the exact location of the recombination event and it may also miss recombination events between markers. In addition, if the two parents carry the same set of alleles, no recombination event between the markers can be identified. Therefore, selection of markers directly affects the effectiveness of re-



(a) Markers on 1st parent.

(b) Markers on 2nd parent.

(c) Markers on the child after recombination.

(d) Markers on the child possibly without recombination.

Figure 1. Identifying a recombination event with markers. The first marker has two alleles A and a. The second marker has two alleles B and b.

combination identification. Typically, a good marker collection should be abundant and evenly distributed across the genome. One such marker family is microsatellites, which are sequences in which a short motif is repeated in tandem [1]. The motifs can be di-, tri-, or tetranucleotide repeat units. There are about 10^4 copies of them distributed quite evenly over the whole genome. They are also shorter and easy to apply PCR. In the Marshfield map [4], over 8000 microsatellites are used; In the Iceland map [2], about 5000 microsatellites are used.

The frequency of recombination varies across the genome rather than uniformly distributed. In almost all chromosomes, recombination is more frequent near the *telomere* – the end of a eukaryotic chromosome; it is often less frequent at the *centromere* where two copies of the homologous chromosomes hold together. Each chromosome is a linear structure of DNAs. We consider X, the location of a recombination event, a random variable. Let $p(x)$ be the p.d.f. of X. Let $F(x)$ be the cumulative distribution function (c.d.f.) of X. The RRD function $R(x)$ is in proportion to $p(x)$, that is:

$$R(x) = R_0 p(x) \tag{1}$$

where R_0 is the total amount of recombination events observed for one individual. According to the definition of $F(x)$, we can also calculate $R(x)$ by:

$$R(x) = R_0 \frac{dF(x)}{dx} \tag{2}$$

In the recent Iceland RRD estimation [2], they used Eq. (1). Since its exact physical location is unknown, a recombination event between two markers is assigned the position of the marker with larger coordinate on the chromosome. With N recombination event locations observed, i.e., x_1, x_2, $\cdots$, x_N, a p.d.f. estimation $\hat{p}(x)$ is obtained

using the Parzen window method, that is:

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} k(x, x_i) \quad (3)$$

where

$$k(x, x_i) = \begin{cases} \frac{1}{\Delta}, & |x - x_i| \leq \frac{\Delta}{2} \\ 0, & \text{otherwise} \end{cases}$$

and Δ is the window width or bandwidth. Then they choose a sequence of M equally spaced locations $y_0, 2y_0, 3y_0, \cdots, My_0$ to calculate the estimated p.d.f. values. In the end, they fit splines to these points to obtain a smooth p.d.f. The RRD is finally obtained by Eq. (1). The critical window width parameter Δ is 3 Mbps. The sample is drawn from 1257 meioses.

A previously published Marshfield RRD [4] used Eq. (2). In this approach, it is not necessary to know the exact location of each recombination event. They first compute the empirical c.d.f. $\hat{F}(x)$ from the observed recombination events, then fit cubic splines to $\hat{F}(x)$. They finally obtain the RRD function $R(x)$ by Eq. (2). In this study, only 184 meioses are analyzed to identify recombination events, which is a much smaller sample size compared to [2].

3. Estimating RRD by Optimal Quantization

Each RRD in [2] is a continuous function. They did not explain how the bandwidth is chosen in the Parzen window approach. In addition, all the splines have to be saved and evaluated before the rate at a certain location can be calculated. Optimal quantization, also a non-parametric technique, is an alternative to their approach. As we shall see, the advantages include that optimal quantization is both statistically and computationally efficient; it is as easy to use as table look-up; over-/under-fitting can be controlled in a systematic way. Optimal quantization finds the most effective representation of data in terms of both CPU cycles, the memory requirement and the targeted performance. Intuitively, an optimal quantization algorithm locates the most important regions which are then finely quantized, while less important regions are coarsely quantized. Importance determination relies on the pattern recognition task. It could be average log likelihood, entropy, or some combination of them. Other methods, e.g., kernel methods, treat everywhere in the space equally without the prioritized resource allocation. For the less important regions, there is the potential wasting of resources.

Our methodology obtains a variable bin width p.d.f. estimation by optimizing a quantizer measure defined on a finite sample, where all bins will have non-zero p.d.f. values. The measure combines convexly average log likelihood and entropy. We perform two steps to find a p.d.f. Dynamic programming is employed to find a quantization of the real axis maximizing the quantizer measure. ensuring the adaptivity to data and overcoming the statistical inefficiency of an equal bin width histogram. The second step obtains a locally averaged p.d.f. estimate for each bin. We use a computationally efficient algorithm for smoothing, which guarantees the consistency of the p.d.f. estimates while avoiding the computation complexity of a kernel type smoothing method. The p.d.f. thus produced is optimal in both the adaptivity and consistency senses.

There are four parameters in our optimal quantization framework: W_J, the weight of log likelihood, W_H, the weight of entropy, number of quantization levels L, number of neighbors k. W_J, W_H and L control the over-/under-fitting in finding the optimal intervals; k controls the over-/under-fitting in estimating the p.d.f. value of each bin. We estimate these parameters by 5-fold cross-validation on the training data. This method is computationally intensive but it can find a reasonably good set of parameters in the spectrum from under- to over-fitting.

4. Experimental Results

We perform optimal quantization on the data from [2]. The genetic distances of the markers are given, corresponding to the empirical c.d.f. of the recombination events. In the optimal quantization experiment, we first obtain the control parameters W_J, W_H, L and k by a 5-fold cross-validation. The values of W_J and W_H range from 0 to 1 with a step of 0.1. The value of L ranges from 2 to 2^8 in the power of 2. The value of k ranges from 1 to 3^6 in the power of 3. Second, the p.d.f. is estimated, using the best parameters just obtained, on all the recombination events of each chromosome. The estimated RRD functions of chromosomes 3, 13, 22 and X are shown in Fig. 2 to 5. Recombination is much more active around the ends of chromosomes than the centers. Our RRD's change more drastically than the ones shown in [2, 4]. Since all our control parameters are cross-validated, it is very likely that the recombination rate changes indeed more abruptly than the much more smooth curves published before. To fit splines on our estimation result could make the curve smoother, but it requires further validation of the smoothness. We further compare quantitatively the performance of optimal quantization with Parzen window approach. We did not apply splines to make the comparison fair. The evaluation is done by a 5-fold cross-validation. The performance measure is the log likelihood of the left-out data reserved for test, using the p.d.f. estimated from the the data without the left-out data. The average and the standard deviation of the five log likelihoods for each chromosome are shown in Table 1. The average log likelihoods of the p.d.f. obtained by optimal quantization are consistently higher than those by Parzen window method. The standard deviations of both are sim-

ilar, with Parzen window results slightly better on most of the chromosomes. Therefore the optimization quantization approach provides a better RRD estimation than that of the Parzen window.

Table 1. Performance comparison between optimal quantization and Parzen window.

Chromosome	average log likelihood		standard deviation	
	Opt. Quant.	Parzen Window	Opt. Quant.	Parzen Window
1	-19.11	-19.17	0.03	0.01
2	-19.10	-19.21	0.05	0.02
3	-18.90	-19.05	0.04	0.04
4	-18.91	-18.98	0.03	0.02
5	-18.79	-18.91	0.04	0.03
6	-18.72	-18.88	0.05	0.03
7	-18.69	-18.87	0.03	0.02
8	-18.60	-18.78	0.02	0.01
9	-18.42	-18.52	0.04	0.03
10	-18.55	-18.69	0.05	0.05
11	-18.53	-18.65	0.06	0.03
12	-18.57	-18.63	0.03	0.04
13	-18.02	-18.32	0.06	0.04
14	-17.94	-18.14	0.07	0.07
15	-17.87	-18.17	0.06	0.07
16	-18.05	-18.18	0.07	0.04
17	-17.99	-18.14	0.05	0.05
18	-18.04	-18.16	0.08	0.06
19	-17.70	-17.95	0.09	0.05
20	-17.62	-17.70	0.09	0.03
21	-17.05	-17.28	0.06	0.05
22	-16.96	-17.16	0.08	0.05
X	-18.42	-18.53	0.04	0.03

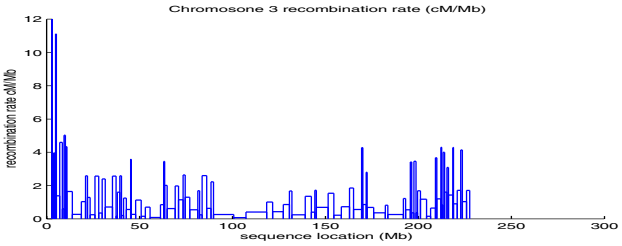


Figure 2. Chromosome 3

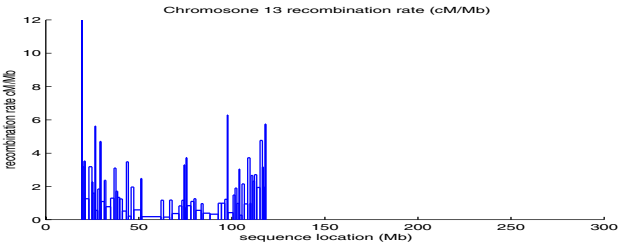


Figure 3. Chromosome 13

5. Conclusion and Future Work

Accurate estimation of the RRD across the entire genome is crucial to understand evolution quantitatively. For example, recombination rate has been known to be linearly correlated with the locations of retrotransposable elements such as long interspersed elements. In contrast to the Parzen window approach used in previously reported RRD's, we have used an optimal quantization approach to obtain piecewise constant RRD's. The optimal partition of each chromosome is obtained first and then the recombination rate in each bin is computed by a neighborhood averaging. Our 5-fold cross-validation result has shown that the performance of optimal quantization in terms of log likelihood is better than the Parzen window approach. The quantization approach produces RRD's that are much easier to use by table look-up than the continuous curves produced by Parzen window and splines. Quantization is an attractive approach to represent the genome-wide distributions of biological events, because of the controled over-/under-fitting and the convenience of use. We further plan to obtain Guanine + Cytosine (GC) percentage distribution by this approach. Currently GC percentages are given by 20Kb windows which obviously is not the most efficient way to do the whole genome of 3 Gb. An even further goal is to use a multidimensional quantization approach to represent different biological features and to look at statistically significant relations revealed by the representation.

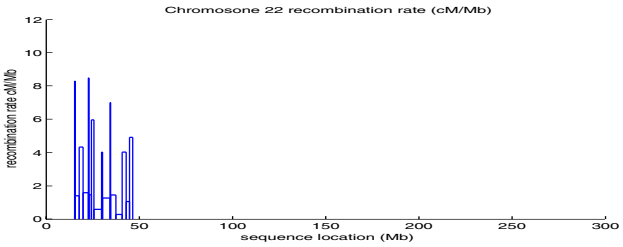


Figure 4. Chromosome 22.

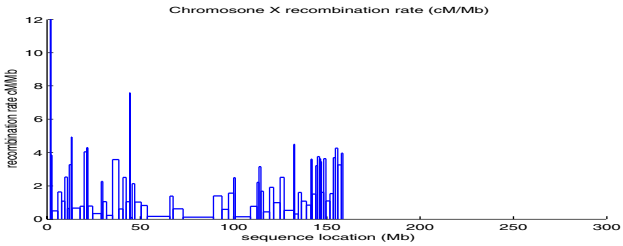


Figure 5. Chromosome X.

References

[1] T. A. Brown. *Genomes*. Wiley-Liss, 1999.

[2] A. Kong and et al. A high-resolution recombination map of the human genome. *Nature Genetics*, 31:241–247, July 2002.

[3] M. Song and R. M. Haralick. Optimally quantized and smoothed histograms. In *Proc. of Joint Conf. of Information Sciences*, pages 894–897, Durham, NC, 2002.

[4] A. Yu and et al. Comparison of human genetic and sequence-based physical maps. *Nature*, 409:951–953, Feb. 2001.