Maximum Likelihood Quantization of Genomic Features using Dynamic Programming

Mingzhou (Joe) Song
Dept. of Computer Science
New Mexico State Univ.
Las Cruces, NM 88003
joemsong@cs.nmsu.edu

Robert M. Haralick
Ph.D. Program in Computer Science
City Univ. of New York
New York, NY 10016
haralick@gc.cuny.edu

Stéphane Boissinot
Department of Biology
Queens College
Flushing, NY 11367
stephane.boissinot@qc.cuny.edu

Abstract

Dynamic programming is introduced to quantize a continuous random variable into a discrete random variable. Quantization is often useful before statistical analysis or reconstruction of large network models among multiple random variables. The quantization, through dynamic programming, finds the optimal discrete representation of the original probability density function of a random variable by maximizing the likelihood for the observed data. This algorithm is highly applicable to study genomic features such as the recombination rate across the chromosomes and the statistical properties of non-coding elements such as LINE1. In particular, the recombination rate obtained by quantization is studied for LINE1 elements that are grouped also using quantization by length. The exact and density-preserving quantization approach provides an alternative superior to the inexact and distance-based k-means clustering algorithm for discretization of a single variable.

1 Introduction

We address the problem of quantization of continuous random variables to discrete ones that will maximally preserve the original probability density function (p.d.f.). We present a method that achieves density-preserving quantization by dynamic programming, which guarantees the optimality of the discretization. In optimal quantization, the most important regions are finely quantized, while less important regions are coarsely quantized, statistically much more efficient than a uniform quantization. Our algorithm can work on either raw continuous data or data already accumulated in finer bins. The number of quantization levels is determined by the Bayesian information criterion – a function of the log likelihood, the sample size, and the number of quantization levels, or cross validation.

Graphical models have motivated continued research on quantization algorithms. A graphical model uses a graph to represent the joint probability distribution function of multiple random variables. Each node in the graph represents a random variable. Edges between nodes encode statistical dependencies among variables. The joint probability distribution function can be decomposed to the product of conditional probability functions of variables at each node given their parents. One strategy to make computation of graphical models feasible is to quantize each variable and treat as multinomial discrete random variables. For discretized random variables, there are more alternatives [1] to determine statistical independencies between them than in the continuous domain when the underlying p.d.f. is unknown. It is often necessary for a quantization to preserve the p.d.f. as much as possible so that to maintain the same statistical dependencies among the discretized variables.

More relevant to our work are approaches that find a quantization of the data by optimizing an quantization objective function. Entropy [2], likelihood[3], and distance have been used as quantization objective functions. Among these criterions, only likelihood ties directly to the p.d.f. of the original continuous random variable. Dynamic programming has provided a less-known optimal solution to the 1-D k-means problem[4], and is later used to find an optimal quantization to classify 1-D samples[5]. However, dynamic programming has not been used in maximizing the likelihood for quantization. Our methodology obtains a non-uniform quantization by optimizing an objective function that combines likelihood and entropy. Optimal quantization ensures the adaptivity to the data and overcomes the statistical ineffectiveness of uniform quantization.

Our quantization algorithm is highly applicable to study genomic features such as the recombination rate across the chromosomes and the statistical properties of non-coding elements such as LINE1. In particular, the recombination rate obtained by quantization is studied for LINE1 elements

that are grouped also using quantization by length.

2 The Likelihood of Quantization

We define and justify a quantization objective function that includes the likelihood and entropy measures on the observed data set. Let X be a random variable and $\mathcal{X}_N = \langle x_1, x_2, \ldots, x_N \rangle$ be a sorted sequence of N real number samples from X, where $x_1 \leq x_2 \leq \ldots \leq x_N$. Let Q be a quantization with L bins. Let $\Delta(q)$ be the width of bin q. Let N_q be the total number of data in bin q. The non-negative Kullback-Leibler divergence from $\hat{p}(x)$ to $p(x)$ is

$$D(p||\hat{p}) = \int p(x) \log \frac{p(x)}{\hat{p}(x)} dx = \mathbf{E}[\log p(X)] - \mathbf{E}[\log \hat{p}(X)],$$

which is to be minimized by a good estimate $\hat{p}(x)$. As $p(x)$ is fixed, maximizing $\mathbf{E}[\log \hat{p}(X)]$ is equivalent to minimizing $D_{KL}(p||\hat{p})$. Let $p(q)$ be the density of bin q. We estimate $\mathbf{E}[\log \hat{p}(X)]$ by average sample log likelihood. Thus the *log likelihood* of X for quantization Q is

$$J(X|Q) = \mathbf{E}[\log \hat{p}(X)] = \frac{1}{N} \sum_{q=1}^{L} N_q \log\left(p(q)\right) = \sum_{q=1}^{L} J(X|q).$$

In [6], entropy is utilized as a class impurity measure, while we use entropy to characterize the generalization ability of quantization. Maximizing entropy corresponds to minimizing information loss. Entropy is defined by

$$H(X|Q) = -\sum_{q=1}^{L} \frac{N_q}{N} \log \frac{N_q}{N} = \sum_{q=1}^{L} H(X|q), \quad (1)$$

where the contribution of a single bin is $H(X|q) = \frac{N_q}{N} \log \frac{N}{N_q}$. Examples of maximum entropy quantization include equal probability quantization [7], histogram equalization [8], Voronoi tessellation [9], or more generally, nearest neighbor partitions [10].

In contrast to likelihood, entropy is not a direct performance measure of pattern recognition test results. Rather, the entropy measure in our context controls over-fitting. The larger the entropy, the less likely for over quantization.

We define the quantization objective function or performance measure as

$$T(X|Q) = w_J J(X|Q) + w_H H(X|Q) \quad (2)$$

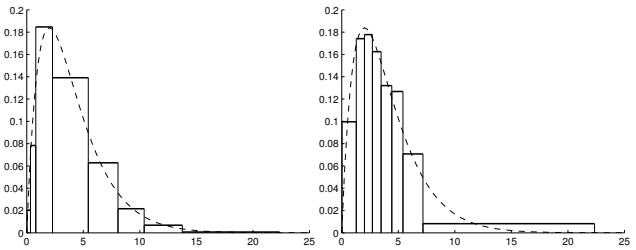
with $w_J + w_H = 1, w_J, w_H \geq 0$, where w_J and w_H are given weights for log likelihood and entropy, respectively. This first term will allow better fit to the data while the second term penalize the complexity of the fitting to avoid over-fitting. $T(X|Q)$ can be written in an additive form as

$$T(X|Q) = \sum_{q=1}^{L} T(X|q) = \sum_{q=1}^{L} w_J J(X|q) + \sum_{q=1}^{L} w_H H(X|q),$$

if we define $T(X|q)$, the contribution of an individual bin q, as $T(X|q) = w_J J(X|q) + w_H H(X|q)$.

A data-driven strategy is to determine the coefficients w_J, w_H through cross validation. The values of w_J, w_H that maximize the likelihood of the left-out fold are selected to be the coefficients.

Example – A Chi-squared example contrasts maximum likelihood and maximum entropy quantization. The 1000 data points were generated using a Chi-squared distribution with 4 degrees of freedom. The quantization level is 8. The density estimates are shown in Fig. 1. The dashed line is the p.d.f. of the Chi-squared distribution. In Fig. 1(a), it



(a) Maximum likelihood quantization ($w_J = 1, w_H = 0$). (b) Maximum entropy quantization ($w_J = 0, w_H = 1$).

Figure 1. Density estimates of Chi-squared data using optimal quantization.

is evident that the underlying density changes much more rapidly in $[0, 2]$ than in $[2, \infty)$. The bins are narrower for the region from 0 to 2 than for the region above 2, corroborating the consistency result in [11]. In Fig. 1(b), the bins for the region around the mode at 2 are narrower than the region further away from the mode. The density of the region around the mode is larger than other regions. When entropy is maximized, each bin contains about the same number of points. This naturally leads to narrower bins for regions of higher density and wider bins for regions of lower density. The rational behind the entropy measure is that the least commitment should be made to the sample. This controls the generalization ability of the quantization. On the other hand, the maximum likelihood approach is always trying to find the best fit to the data and it may over-fit. Therefore, it is necessary to combine the two measures in a controlled fashion as we have done in defining $T(X|Q)$.

3 The Optimality Condition for Quantization using Dynamic Programming

We state without proof the optimality condition for solving the quantization problem using dynamic programming. Given X and the number of quantization levels L, the goal of quantization is to find decision boundaries $B =$

$\{b_0, b_1, \ldots, b_L\}$, $b_0 < b_1 < \ldots < b_L$, such that a predefined objective function $F(X, B)$ is maximized. The optimality condition is that if there exists a monotonically increasing function $g(x)$ such that

$$g(F(X, B)) = \sum_{q=1}^{L} g(F(X_q, B_q)), \tag{3}$$

where $B_q = (b_{q-1}, b_q]$ and $X_q = \langle x | x \in B_q \rangle$, then finding a quantization to maximize $F(X, B)$ can be solved exactly using dynamic programming.

4 Maximum Likelihood Quantization using Dynamic Programming

Evidently the optimality condition Eq. (3) holds for $T(X|Q)$ with $g(x) = x$. Thus, we can use dynamic programming to find a quantization that maximizes $T(X|Q)$. To avoid over-fitting, one can require a minimum number of k data in each bin and that identical data are put into the same bin. We only set the decision boundaries in the middle of two consecutive data points. This affects the range of $J(X|Q)$, but it is trivial when sample size is not too small. This restriction prevents $J(X|Q)$ from overflow.

We define the probability density in bin q by $p(q) = \frac{N_q/N}{\Delta(q)}$. We also define the performance measure of a sub-quantization Q_r^u by $T(X|Q_r^u) = \sum_{q=r}^{u} T(X|q)$. We use the notation $T(X|Q_r^u, \mathcal{X}_m^n) = \sum_{q=r}^{u} T(X|q)$ to indicate that the sub-quantization is to be evaluated on the data set $\mathcal{X}_m^n$ that falls in its bins. Note N is still defined on the overall data set $\mathcal{X}_N$, not $\mathcal{X}_m^n$. Let $T[n, q]$ be the maximum performance measure from bin 1 to q when x_n is the largest data in bin q. Let $I[n, q]$ be the index to the smallest element in bin q such that $T[n, q]$ is achieved. Let $T^1[i, n]$ be the performance measure contributed by a bin containing exactly x_i to x_n. The dynamic programming to maximize $T[N, L]$ is described below.

Initialization – $T[0, 0] = I[0, 0] = 0$, $I[0, q] = -1$ for $q \in \{1, \cdots, L\}$, $I[n, 0] = -1$ for $n \in \{1, \cdots, N\}$, $I[n, q] = -1$ for $(n, q) \in \{(n, q) | 0 \leq q < \max(1, n - (N - L))$ or $\min(n, L) < q \leq L, 1 \leq n \leq N, 1 \leq q \leq L\}$.

Feasible decision boundary index set – The indices of the feasible data for being the smallest element in bin q form the feasible decision boundary index set $\mathcal{A}_q^n = \{i |, i \leq n - k + 1, I[i - 1, q - 1] \neq -1, x_{i-1} \neq x_n, I[n, q] \neq -1, x_n \neq x_{n+1}\}$. The inequality $i \leq n - k + 1$ guarantees at least k data in bin q. $I[i - 1, q - 1] \neq -1$ states that x_{i-1} must be feasible for the largest element in the previous bin $q - 1$. $x_{i-1} \neq x_n$ enforces that the feasible largest element in the previous bin $q-1$ must not be the same as x_n, to avoid splitting equally valued data into different bins. $x_n \neq x_{n+1}$ is also not to split equally valued data. $I[n, q] \neq -1$ asserts that x_n must be feasible for the largest element of bin q.

Recurrence – If $\mathcal{A}_q^n$ is empty, then $I[n, q] \triangleq -1$, meaning x_n does not qualify for the largest element in bin q. Otherwise,

$$T[n, q] \triangleq \max_{i \in \mathcal{A}_q^n} T[i - 1, q - 1] + T^1[i, n], \tag{4}$$

$$I[n, q] \triangleq \operatorname*{argmax}_{i \in \mathcal{A}_q^n} T[i - 1, q - 1] + T^1[i, n]. \tag{5}$$

We assert that $T[N, L]$ maximizes the performance measure, the corresponding partition is an optimal solution, and the algorithm has time complexity $O(LN^2)$. The weights w_J and w_H can be determined by cross-validation. The choice of L can be made using either Bayesian information criterion or cross validation. The dynamic programming working with sample points can be readily changed to apply to cumulative sample or data that are already binned, because the algorithm uses only counts of data within an interval rather than the actual values of those points.

5 Estimation of Recombination Rate Distribution over Chromosomes by Quantization

Recombination is central to molecular evolution. The mechanism of recombination can reveal directly how human evolution occurs. In the nucleus of each human cell except the gamete, each of chromosomes 1 to 22 comes with two copies of autosomes called *homologous chromosomes*. During meiosis, the chromosomes of a child emerge by combining half of the chromosomes from one parent and half from the other parent. Only homologous chromosomes will be combined; the two sex chromosomes always combine themselves. During combination, the contents of the chromosomes are exchanged at some points along the chromosomes, which could be due to cross-over or gene conversion. Thus child chromosomes do not necessarily contain exactly same copies of parent chromosomes. This information exchange between parent chromosomes is called *recombination*. *Recombination rate* (RR) 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 RR distribution (RRD) function maps a location on the chromosome to an RR value. However, experimental data on recombination are still very limited due to the cost of experiments. As the complete human genome physical map becomes available, an accurate quantitative representation of the RRD becomes possible.

Recombination events are identified using both genetic and physical maps. On a genetic map, each marker represents a unique feature. A marker has two or multiple forms, called *alleles*. The alleles can be identified by polymerase chain reaction. Locations of markers on the physical map are determined in advance. Markers make detection of recombination events possible without sequencing the entire

genomes of generations. The resolution of the identified events increases with the number of markers. This method is illustrated in Fig. 2. The first parent has 2 markers A and B (Fig. 2(a)) and the second parent has the same markers but with different alleles a and b (Fig. 2(b)). If the child has the markers as in Fig. 2(c), then at least one recombination event has occurred at some location between markers A and B. If the child has the same alleles as their parents as in Fig. 2(d), then it is unlikely to have a recombination event between A and B if the markers are close enough. This

----A-----B---- ----a-----b----

(a) Markers on 1st parent. (b) Markers on 2nd parent.

----A-----b---- or ----a-----B----

(c) Markers on the child after recombination.

----A-----B---- or ----a-----b----

(d) Markers on the child possibly without recombination.

Figure 2. Identifying a recombination event with markers. One marker has two alleles A and a; the other has two alleles B and b.

method cannot detect the exact location of a recombination event or it may miss a recombination event between markers. In addition, if the two parents carry the same set of alleles, no recombination event between the markers may be identified. Therefore, selection of markers directly affects the effectiveness of recombination detection. Typically, a good marker collection should be abundant and evenly distributed across the genome. One such marker family is microsatellites, which are short sequences of motifs in tandem [12]. The motifs can be di-, tri-, or tetra-nucleotide repeat units. There are about 10^4 copies of them distributed quite evenly over the genome. In the Marshfield map [13], over 8,000 microsatellites are used; In the Iceland map [14], there are 5,000 microsatellites.

The frequency of recombination is not uniform across the genome: more frequent near the *telomere* – the end of a eukaryotic chromosome – and less frequent at the *centromere* where two copies of the homologous chromosomes hold together. We consider X, the location of a recombination event, a random variable. Let $p(x)$ be its p.d.f. Let $F(x)$ be its cumulative distribution function (c.d.f.).

The RRD function $R(x)$ is in proportion to $p(x)$ defined as $R(x) = R_0 p(x)$, where R_0 is the total amount of recombination events observed on a single chromosome of an individual. This definition is used in the Iceland RRD estimation [14]. 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 in [14]

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} k(x, x_i), \tag{6}$$

where

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

and Δ is the 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 $p(x)$ and then obtain $R(x)$. The critical bandwidth parameter Δ is 3 Mbps. The sample is drawn from 1257 meioses.

Another RRD is defined in $F(x)$ by $R(x) = R_0 \frac{dF(x)}{dx}$, used by the Marshfield RRD [13]. In this approach, it is not necessary to know the exact location of each recombination event. They compute the empirical c.d.f. $\hat{F}(x)$ from the observed recombination events, then fit cubic splines to $\hat{F}(x)$ and then obtain the RRD. In this study, only 184 meioses are analyzed to identify recombination events, which is a much smaller sample size compared to [14].

The RRDs in [14] are represented as continuous functions, with no explanation of how the bandwidth Δ is chosen. All the splines are saved and must be evaluated to calculate RRD at a location. Alternatively, an optimal quantization algorithm locates the most important regions which are then finely quantized, while less important regions are coarsely quantized. Other methods, e.g., kernel methods, treat everywhere in the space equally without prioritized resource allocation. In the less important regions, the potential of waste of resources exists.

We performed optimal quantization on the genetic distances of selected markers [14], given as the empirical c.d.f. of the recombination events. We first obtained 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. L ranges from 2 to 2^8 in powers of 2. k ranges from 1 to 3^6 in powers of 3. Second, using the best parameters, the p.d.f. was estimated, on all the recombination events of each chromosome. The estimated RRD functions of chromosomes 3 and X are shown in Fig. 3 and 4. Recombination is much more active around the ends of chromosomes than the centers. Our RRDs show more fluctuations than those shown in [14, 13]. Since our control parameters are all cross-validated, it is very likely that the RRDs indeed change 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 validation of the smoothness. We further compare quantitatively the performance of optimal quantization with Parzen window approach. To make the comparison fair, we did not apply splines. 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 not using the left-out data. The average and the standard deviation of the cross-validated log likelihood 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 similar, with Parzen window results slightly smaller on most of the chromosomes. Therefore the optimization quantization approach provides a better RRD estimation than that of the Parzen window.

Table 1. Comparison between optimal quantization & 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

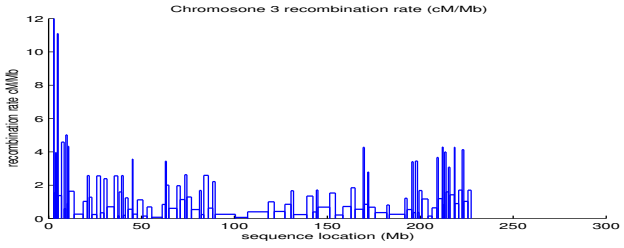


Figure 3. Chromosome 3

6 Localized Study of Recombination Rate within Length Groups of LINE1s

The abundance of LINE-1 (L1) retrotransposons constitutes one of the most puzzling features of mammalian genomes and it is now clear that L1 retrotransposons have

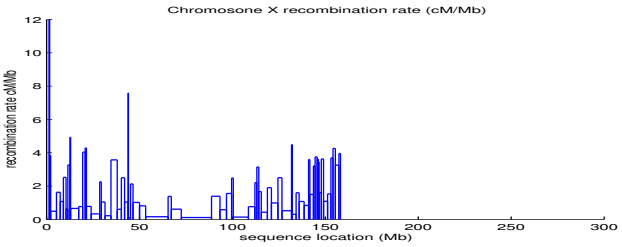


Figure 4. Chromosome X.

profoundly affected the structure and function of genomes [15, 16, 17]. However, the evolutionary forces underlying their genomic distribution and their dynamics in natural populations remain incompletely understood. Although L1 insertions can occasionally be recruited to perform a function beneficial to the host [17], the vast majority of new insertions are more likely to be either neutral or detrimental.

One interesting question is how the RR at a L1 location might depend upon the length of the L1. A linear regression could not adequately capture subtlety of the RR-length interaction. Given the relatively large sample size of L1s, instead of fitting a higher order linear regression model, we broke L1s in L1PA2 to L1PA6 families into groups by their length and looked at the trend of RR within each group. Grouping is determined by optimal quantization of the lengths of all L1s under consideration. Intuitively, this method separates L1s into groups by length when there is a sudden change in the number of L1s over unit length. We selected the number of groups to be six, roughly capturing the overall distribution of length while assuring that the intervals are not too small for a meaningful regression. The six length groups are shown in Table 2. The grouping reflects a natural tendency for L1 to segregate by length.

Table 2. L1 groups by length, with length ranges, counts, and percentage.

L1 Groups	Length Range	L1 Count/Percentage
1	[100,490]	12226/34%
2	[491,1152]	8559/24%
3	[1153,2498]	6462/18%
4	[2499,6001]	4182/12%
5	[6002,6183]	4231/12%
6	$\geq$ 6184	218/1%

A one-way ANOVA (Table 3) indicates indeed the RR means are significantly different among L1 length groups. The Tukey's Honest Significant Differences (HSD) test reveals further details in Fig. 5. Under the null hypothesis of RR mean equality across groups, if one compares every

Table 3. One-way ANOVA for RR over the length groups.

	Degrees of Freedom	Sum of Squares	Mean Squares	F value	Pr($> F$)
group	5	93	19	7.5441	4.330e-07
Residuals	35872	88107	2		

two groups using the 5% α-level, the chance of observing some inequality among the pairs can be much greater than the anticipated 5% type I error. The Tukey's HSD test corrects this problem. In Fig. 5, the range of each line segment manifests the 95% confidence interval of the mean RR difference between the two length groups labeled on the left of the segment. The vertical dashed line marks the zero difference location. If an interval contains zero, there is no significant evidence from the sample to conclude that the two groups have different mean RRs. All differences are the mean RR of a group with a longer length minus that of one with a shorter length. A major observation is that no segments have both ends above zero, suggesting no significant trend of increasing RR as length increases. The only almost significant negative difference between two consecutive length groups occurs from group 2 to 3, which accounts for other significant differences among non-consecutive length groups. Therefore, the multiple comparison analysis pins down that the most significant reduction in RR takes place among the L1s of intermediate length.

Based on the Tukey's HSD results, we studied the trend of RR within each length group using linear regression on the length of L1. The intercepts and slopes of each linear regression line, and the corresponding p-values are given in Table 4. No length group shows a significant positive slope. We observe that the len.cat2 group has a highly significant negative slope. Figure 6 shows the mean RR-length scatter

Table 4. Linear regression slopes of each group.

	Estimate	Std. Error	t-Statistic	Pr($> \|t\|$)
1:length	-5.537e-05	1.275e-04	-0.434	0.6641
2:length	-2.446e-04	9.006e-05	-2.716	0.0066
3:length	-3.409e-05	5.126e-05	-0.665	0.5060
4:length	3.042e-06	2.268e-05	0.134	0.8933
5:length	5.923e-05	4.409e-04	0.134	0.8931
6:length	3.108e-04	5.386e-04	0.577	0.5639

plot with the regression lines overlaid. We can observe in the plot a decreasing trend of the regression line in group

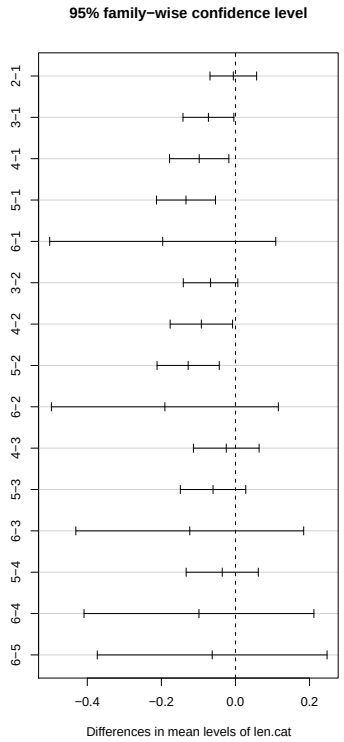


Figure 5. Tukey's HSD test on the RR means among length groups. Numbers on the vertical axes correspond to length groups. For example, 5-3 stands for the mean RR of group 5 minus that of group 3.

2 quite evidently. It is also quite evident subjectively that there is a declining tendency in the mean RR as the length increases. This further analysis match well to previous findings by the Tukey's HSD test. Therefore the major RR reduction occurs on the L1s of length 491 to 1152, which are not full-length L1s, but the L1s of intermediate length.

7 Conclusion

We have described an approach to quantize optimally a random variable based on likelihood by dynamic programming. Although our approach is quadratic in sample size, it guarantees the optimality. The distance-based k-means algorithm for 1-D quantization due to its computational convenience shall either be replaced by our likelihood-based approach when preservation of the distribution of the original continuous random variables is desired, or by a dynamic programming implementation similar to ours that guarantees optimality. Applications of our algorithm in estimating RR distributions and characterizing LINE1 elements show

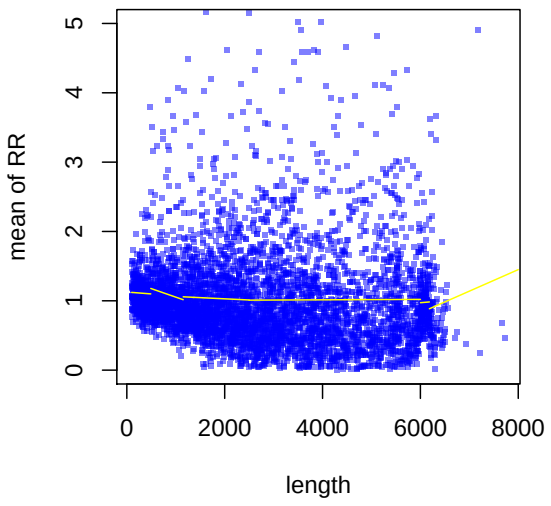


Figure 6. Scatter plot of mean RR versus L1 length. The line segments are linear regressions within each group. Only the second segment has a significant decreasing trend.

its effectiveness in capturing the underlying p.d.f.s of data. It can also be used to discretize other genomic features including GC content, gene expression rate, and non-coding element densities over an entire genome.

Acknowledgment

The authors thank the support from grants made by PSC-CUNY, CUNY Institute for Software Design and Development, and NSF CREST Center for Excellence in Computational Biology and Bioinformatics (Grant Number: HRD_0420407).

References

[1] D. Margaritis and S. Thrun., "A Bayesian multiresolution independence test for continuous variables," in *17th Conference on Uncertainty in Artificial Intelligence (UAI)*, Seattle, Washington, 2001.

[2] R. M. Haralick, "The table look-up rule," *Communications in Statistics – Theory and Methods*, vol. A5, no. 12, pp. 1163–91, 1976.

[3] L. B. Hearne and E. J. Wegman, "Maximum entropy density estimation using random tessellations," in *Computing Science and Statistics*, vol. 24, 1992, pp. 483–7.

[4] X. Wu, "Color quantization by dynamic programming and principal analysis," *ACM Trans. Graph.*, vol. 11, no. 4, pp. 348–372, 1992.

[5] T. Fulton, S. Kasif, and S. L. Salzberg, "Efficient algorithms for finding multi-way splits for decision trees," in *Proc. 12th Int'l Conf. on Machine Learning*, 1995, pp. 244–251.

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Pacific Grove, California: Wadsworth & Brooks/Cole, 1984.

[7] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," vol. SMC-3, no. 6, pp. 610–621, November 1973, see Appendix for equal-probability quantization.

[8] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliff, NJ: Prentice Hall, 1989.

[9] G. Voronoi, "Nouvelles applications des parametres continus a la théorie des formes quadratiques, deuxieme memoire, recherches sur les parallelloedres primitifs," *Journal für die Reine und Angewandte Mathematik*, vol. 134, no. 198-287, 1908.

[10] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

[11] D. W. Scott, *Multivariate Density Estimation – Theory, Practice and Visualization*. John Wiley & Sons, 1992.

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

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

[14] A. Kong and et al, "A high-resolution recombination map of the human genome," *Nature Genetics*, vol. 31, pp. 241–247, Jul. 2002.

[15] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, and et al, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[16] R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, and et al, "Initial sequencing and comparative analysis of the mouse genome," *Nature*, vol. 420, pp. 520–562, 2002.

[17] H. H. Kazazian, "Mobile elements: drivers of genome evolution," *Science*, vol. 303, pp. 1626–1632, 2004.