AN MDL METHOD FOR FINDING HAPLOTYPE BLOCKS AND FOR ESTIMATING THE STRENGTH OF HAPLOTYPE BLOCK BOUNDARIES

M. KOIVISTO,[1] M. PEROLA,[2,3] T. VARILO,[3] W. HENNAH,[3] J. EKELUND,[3]
M. LUKK,[1] L. PELTONEN,[2,3] E. UKKONEN,[1] H. MANNILA [1]

[1] *Department of Computer Science and HIIT Basic Research Unit, University of Helsinki, Helsinki, Finland;* [2] *Department of Human Genetics, UCLA School of Medicine, Los Angeles, CA;* [3] *Department of Molecular Medicine, National Public Health Institute, Helsinki, Finland.*

We describe a new method for finding haplotype blocks based on the use of the minimum description length principle. We give a rigorous definition of the quality of a segmentation of a genomic region into blocks, and describe a dynamic programming algorithm for finding the optimal segmentation with respect to this measure. We also describe a method for finding the probability of a block boundary for each pair of adjacent markers: this gives a tool for evaluating the significance of each block boundary. We have applied the method to the published data of Daly *et al.*[1] The results are in relatively good agreement with the published results, but also show clear differences in the predicted block boundaries and their strengths. We also give results on the block structure in population isolates.

1 Introduction

Haplotype blocks [1,2,3,4] form fascinating small-scale structure in the human genome. While several studies have confirmed that some type of block structures exist, the recent data about haplotype blocks in the human genome have left multiple uncertainties concerning block boundaries and their variation. The published algorithms by Daly *et al.*,[1] Patil *et al.*,[2] and Zhang *et al.*,[3] have used segmentation algorithms with fairly ad hoc criteria for block quality. Also, the existing methods produce a segmentation without any clear indication on how strong or weak the different block boundaries are.

We describe a new method for finding haplotype blocks based on the use of the minimum description length principle. We give a rigorous definition of the quality of a segmentation of a genomic region into blocks, and describe a dynamic programming algorithm for finding the optimal segmentation with respect to this measure. We also describe a method for finding the probability of a block boundary for each pair of adjacent markers: this gives a tool for evaluating the significance of each block boundary.

Intuitively, a haplotype block can be considered to be a sequence of markers such that for those markers most of the haplotypes in the population fall into a small number of classes; each class consists of identical or almost identical haplotypes. This notion can be formalized by considering the problem of describing the haplotypes in a succinct way. This approach is an instance of the minimum description length principle (MDL) by Rissanen [5,6] widely used in statistics, machine learning, and data mining (see, e.g., Li and Vitanyi [7] or Hansen and Yu [8]). Similar ideas have also been applied to partitioning homogeneous DNA domains.[9]

We have applied the method to the published data of Daly *et al.* [1] The results are in relatively good agreement with the published results, but also show clear differences in the predicted block boundaries and their strengths. The method has also been applied to samples from isolated populations.

The rest of this paper is organized as follows. In Section 2 we describe the MDL principle and the encoding of the data. Section 3 gives the dynamic programming algorithm, and Section 4 shows how dynamic programming can be used to compute the probabilities of block boundaries. Section 5 gives an overview of the empirical results, and Section 6 is a short conclusion.

2 MDL principle and coding of haplotype data

Let D be an $n \times p$ matrix of n observations over p markers. We refer to the jth allele of observation i by D_{ij}. For simplicity, we first assume that $D_{ij} \in \{0, 1\}$.

A marker interval $[a, b] = \{a, a+1, \ldots, b\}$ is defined by two marker indices $a, b \in \{1, \ldots, p\}$. A *segmentation* is a set of non-overlapping non-empty marker intervals. A segmentation is *full* if the union of the intervals is $[1, p]$. The data matrix limited to interval $[a, b]$ is denoted by $D(a, b)$ and the values of the ith observation are denoted by $D(i, a, b)$.

The minimum description length principle by Rissanen [5,6] considers the description of the data using two parts: the model B and the description of the data D given the model. The description length for the data and the model is

$$L(B, D) = L(B) + L(D \mid B),$$

where $L(B)$ is the length of the description of the model and $L(D \mid B)$ is the length of the description of the data, when the data is described using the model B.

The minimum description length principle states that the desired descriptions of the data are ones having the minimum length $L(B, D)$ of the total description. For a good survey of the (sometimes intricate) connections between MDL, Bayesian statistics, and machine learning, see Li and Vitanyi.[7] The MDL principle has successfully been used in various applications.[10,11,12,8]

The haplotype data set D can be described by telling first of all how many blocks there are and where the blocks start and end. For each block, we have to specify how many typical haplotypes (*class center*) there are and what they are. For each observation and each block, we tell from which typical haplotype the observation comes from.

More formally, a block model B consists of the following components.

1. A segmentation S, i.e., the start and end markers s_h and e_h for each block $(h = 1, \ldots, \ell)$.[a] Implicitly, the segmentation specifies the number of blocks ℓ.

2. For each block h, the class centers $\theta_h = (\theta_{hc})$ $(c = 1, \ldots, k_h)$ specifying the coordinates θ_{hcj} for each marker $j = s_h, \ldots, e_h$. Implicitly, each θ_h also specifies the number of centers k_h.

The coordinates θ_{hcj} are real numbers, encoding the probability of seeing 1 in marker j of an observation stemming from class center c of block h. So, strictly speaking, a class center is not a typical haplotype, but a mean vector of the haplotypes associated with the class.

Given such a block model $B = ((s_h, e_h), (\theta_{hcj}))$, the data can be encoded as follows. For each observation $i = 1, \ldots, n$, and for each block $h = 1, \ldots, \ell$, we first have to tell which of the k_h class centers does the observation $D(i, s_h, e_h)$ belong to; let this center be c. This takes $\log k_h$ bits per observation. Then we have to describe $D(i, s_h, e_h)$ using the center coordinates θ_{hcj}, for $j = s_h, \ldots, e_h$. This is done by assuming independence of marker values given the class center. Thus the probability is

$$P(D(i, s_h, e_h) \mid (\theta_{hcj})) = \prod_{j=s_h}^{e_h} \theta_{hcj}^{D_{ij}} (1 - \theta_{hcj})^{1-D_{ij}}. \tag{1}$$

Using the relation of coding lengths and probabilities we get a code of length $-\log P(D(i, s_h, e_h))$ for the data D, given the segmentation model B.

Assuming the segmentation in the block model B is full, it can be coded using $\ell \log p$ bits for encoding the block boundaries, and $\ell \log n$ bits for the number of centers in the block, and $\alpha k_h(e_h - s_h + 1)$ bits for coding the centers, where α is the number of bits needed for the coding of a real number. Theoretical arguments [5,6,8] indicate that the appropriate accuracy is obtained by choosing $\alpha = (\log n)/2$.

Thus the length of the description of the block model is

$$L(B) = \ell \log p + \ell \log n + \sum_{h=1}^{\ell} k_h \alpha (e_h - s_h + 1).$$

[a] Of course, $s_1 = 1$ and $e_\ell = p$. In some of our models we allow parts of the data to be uncoded, so $s_{h+1} = e_h + 1$ does not necessarily hold.

The length of the description of the data is

$$L(D \mid B) = \sum_{h=1}^{\ell} \sum_{i=1}^{n} \big[\log k_h - \log P(D(i, s_h, e_h)) \big].$$

Thus the goal of the segmentation procedure is to find a block model B such that the overall coding length $L(B, D) = L(B) + L(D \mid B)$ is minimized.

The description method is easily extended to handle missing or unknown data values. If the values D_{ij} are interpreted as a degree of certainty that the correct value is 1, the expression in Eq. 1 can be used directly. For instance, one can assign the value 0.5 for each missing allele in the data. To obtain a proper probability model a normalizing factor should be included in Eq. 1. However, the factor behaves as an irrelevant constant and therefore can be ignored.

3 Dynamic programming algorithm

We use a dynamic programming algorithm to compute an optimal block structure, and then estimate the probabilities of each block boundary.

The MDL cost function is, as defined above, a function of the whole segmentation. However, it is straightforward to see that it can be decomposed into the blocks of the segmentation. Given a marker interval $[a, b]$, let $\hat{k}$ is the optimum number of centers, and let $\hat{\theta}_{ij}$ be the corresponding center coordinates associated with jth allele of ith observation, such that the cost

$$\begin{aligned} f(a, b) \quad = \quad & \log p + \log n + n \log \hat{k} + \frac{1}{2} \hat{k} (b - a + 1) \log n \\ & + \sum_{i=1}^{n} \sum_{j=a}^{b} \big[- D_{ij} \log \hat{\theta}_{ij} - (1 - D_{ij}) \log(1 - \hat{\theta}_{ij}) \big] \end{aligned} \qquad (2)$$

is minimized. Then for the MDL optimal block model B_{mdl} we have

$$L(B_{mdl}, D) = \min_{S} \sum_{[a,b] \in S} f(a, b),$$

where S runs through all full segmentations on $[1, p]$. Thus the minimum description length of haplotype data can be defined as the sum of costs of coding of individual blocks.

Denote by $F(b)$ the cost of the optimal segmentation of the haplotypes from marker 1 to marker b. We have the typical dynamic programming equation

$$F(b) = \min_{1 \le a \le b} (F(a - 1) + f(a, b));$$

additionally $F(0)$ is defined to be 0. Namely, the coding from marker 1 to marker b is either produced by coding all the markers in one block (with cost $F(0)+f(1,b)$), or by coding for some a from marker 1 to marker $a-1$ optimally (cost $F(a-1)$) and then coding from marker a to marker b in one block (cost $f(a,b)$). Given the costs $f(a,b)$, the computation can be done in $O(p^2)$ time.

The cost $f(a,b)$ is computed by using k-means clustering on the data set $D(a,b)$. The number of cluster centers is varied from 1 to 10, and for each number we produce 5 different clusterings. For each clustering, the coding cost of $D(a,b)$ is computed, and as the cost $f(a,b)$ we select the smallest cost.[b] The computation of $f(a,b)$ takes time $O(n(b-a+1))$ for a fixed number of iterations in the k-means algorithm. Thus the total amount of time needed for computing the costs $f(a,b)$ is $O(np^3)$.

In many cases it is interesting to see how optimal segmentations behave when some (bad) markers are allowed to be ignored in the data. We call such ignored markers *gaps* between haplotype blocks. A natural extension of the problem of finding the optimum segmentation is to find a segmentation that gives the shortest description length and includes at most u gaps. Denoting by $F(b,u)$ the cost of optimal segmentation from marker 1 to marker b using at most u gaps, we have

$$F(b,u) = \min(F(b-1,u-1), \min_{1\leq a\leq b}(F(a-1,u)+f(a,b))).$$

Namely, if a gap is used at the bth marker, then the prefix segmentation of $[1,b-1]$ is allowed to contain at most $u-1$ gaps. Otherwise, a block $[a,b]$ is introduced and the maximum allowed number of gaps from marker 1 to marker $a-1$ is still u. The computation of $F(b,u)$ can be arranged to take $O(np^3)$ time.

4 Computing the probability of a block boundary

We next consider the probability that there is a block boundary between markers j and $j+1$. Denote by $\mathcal{S}_{j,j+1}$ the set of all full segmentations having a boundary between markers j and $j+1$. Then we are interested in the probability of any segmentation from $\mathcal{S}_{j,j+1}$, given the data D:

$$P(\mathcal{S}_{j,j+1} \mid D) = \sum_{S\in\mathcal{S}_{j,j+1}} P(S \mid D).$$

[b] The problem of finding the best cluster centers is NP-hard; thus the approach does not guarantee that the shortest description for the single block from a to b is found.

Denoting by $\mathcal{S}[1,p]$ the set of all full segmentations on $[1,p]$, this can be written

$$P(\mathcal{S}_{j,j+1} \mid D) = \frac{\sum_{S\in\mathcal{S}_{j,j+1}} P(S,D)}{\sum_{S'\in\mathcal{S}[1,p]} P(S',D)}. \tag{3}$$

The probabilities $P(S,D)$ come naturally from our description method. For any segmentation S and data set D we define

$$P(S,D) = Z^{-1}\, 2^{-\sum_{[a,b]\in S} f(a,b)},$$

where $f(a,b)$ are the minimum description lengths for the corresponding blocks as described in Eq. 2, and Z is a normalization constant that does not depend on S and D. Note that the normalization constant cancels out when substituted into Eq. 3.

Define

$$q(a,b) = 2^{-f(a,b)},$$

and for any interval $[j,j']$,

$$Q(j,j') = \sum_{S\in\mathcal{S}[j,j']} \prod_{[a,b]\in S} q(a,b),$$

where $\mathcal{S}[j,j']$ denotes the set of all full segmentations on marker interval $[j,j']$. Then, since $\mathcal{S}_{j,j+1}$ is equal to the Cartesian product $\mathcal{S}[1,j]\times\mathcal{S}[j+1,p]$, we have

$$P(\mathcal{S}_{j,j+1} \mid D) = \frac{Q(1,j)Q(j+1,p)}{Q(1,p)}.$$

(For a similar development, see Durbin *et al.*,[13] Eq. 3.14, and also Liu and Lawrence.[14])

Again, dynamic programming can be applied. The equations are

$$Q(1,b) = \sum_{1\le a\le b} Q(1,a-1)q(a,b)$$

and

$$Q(a,p) = \sum_{a\le b\le p} q(a,b)Q(b+1,p).$$

(Here, of course, we define $Q(1,0) = Q(p+1,p) = 1$.) Thus the probabilities $P(\mathcal{S}_{j,j+1} \mid D)$ can be computed for all j in time $O(p^2)$.

Table 1: Haplotype classes of the block 2 found by the MDL method for the data of Daly *et al.* The block consists of the markers 15–24. There are two classes of haplotypes. For each class the size (the number of associated haplotypes), and the center coordinates and the most commonly occuring haplotype is shown. The right-most columns show the number of haplotypes that differ from the most commonly occuring haplotype at 0, 1, 2, or ≥ 3 markers.

Class	Size	Center and the most frequent haplotype										Frequencies			
												0	1	2	≥ 3
1	190	.16	.02	96	.64	.06	.15	.05	.02	.94	.02				
		0	0	1	1	0	0	0	0	1	0	115	44	20	11
2	68	.91	.88	.18	.20	.82	.20	.82	.92	.15	.91				
		1	1	0	0	1	0	1	1	0	1	34	18	4	12

5 Empirical results

We have tested the above methods on synthetic (generated) and real data. The results on synthetic data show that the method finds the block structure that have been used to generate the data, and that the method is quite robust against noise (data not shown).

Figure 1 shows the results when the MDL method is applied to the data of Daly *et al.*[1] The segmentations reported by Daly *et al.*[1] and those produced by the MDL method are overall in quite good agreement. There are, however, some differences. For example, the 3 block boundaries around marker 40 reported in[1] are clearly weaker than some of the others, as shown by the log odds curve, and also in the plot of the optimal segmentations with a varying number of gap markers.

An interesting difference can also been found at the block around marker 20. Since Daly *et al.*[1] count exact matches they report 3 distinctive common haplotypes. Our method, however, suggests 2 haplotype classes, for it allows variability within a class, see Table 1.

Figure 2 displays the results when random noise is added to the data of Daly *et al.* The block structure as well as the probabilities for block boundaries show quite good stability. As expected, when the marker order is randomly permuted, the block structure disappears. Note that some information already shown in Figure 1 is repeated, but using the physical location of the markers yields additional insights into the nature of the blocks.

We also applied the method to SNP data on samples from three subpopulations from Finland, which are representative for the settlement history of Finland, inhabited by two periods of immigration 4000 and 2000 years ago.[15] The early settlement sample ($n = 32$ chromosomes) consisted of descendants of the early settlers ($\approx$100 generations) on the coastal areas of Finland, the late

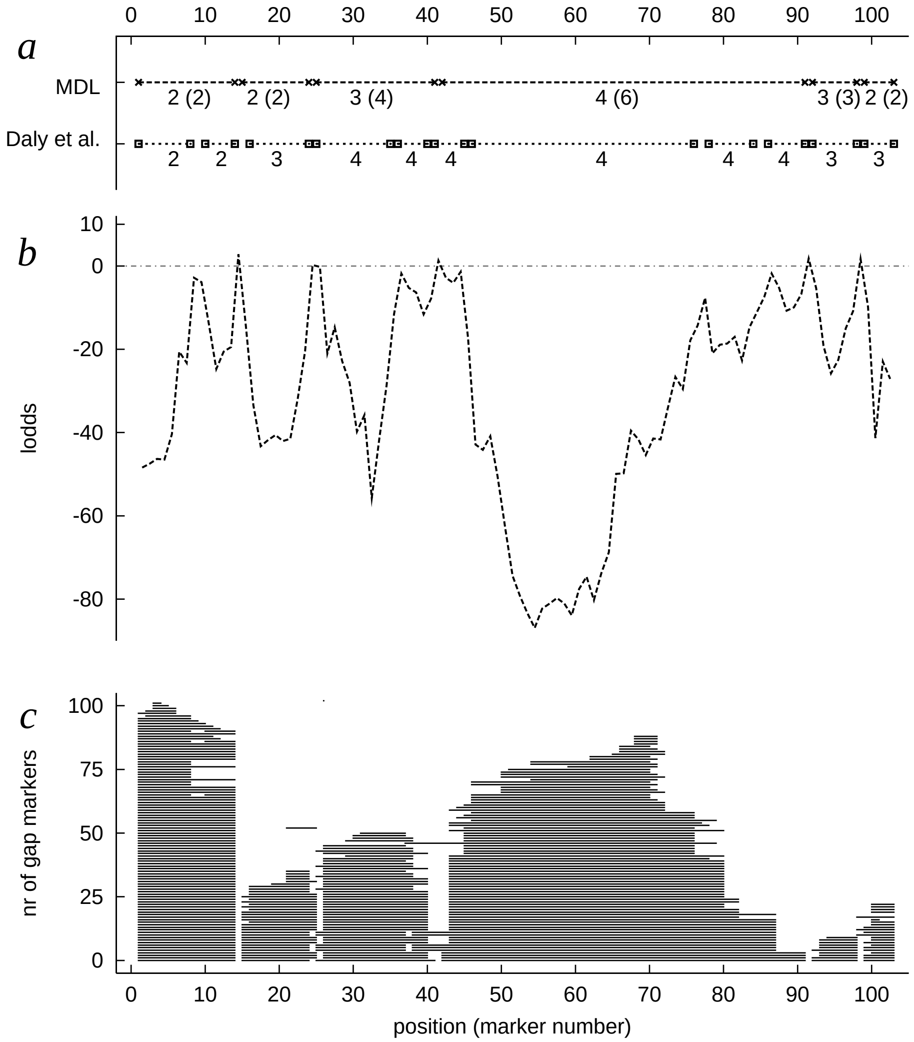

Figure 1: Haplotype block structure in the data of Daly *et al.* The x-axis is the number of the marker. (a) The optimal block structure produced by the MDL scoring function compared against the block boundaries reported by Daly *et al.* The numbers associated with the MDL blocks give the number of haplotype classes that suffice to cover at least 85 percent of the block and, in parenthesis, the total number of the classes. The numbers associated with the Daly *et al.* blocks give the number of haplotypes in the block that suffice to cover at least 90 percent of the block. (b) The log odds of the probability of block boundaries for each pair of adjacent markers. (c) The optimal segmentation when k markers are allowed to be left outside the blocks, for varying k.

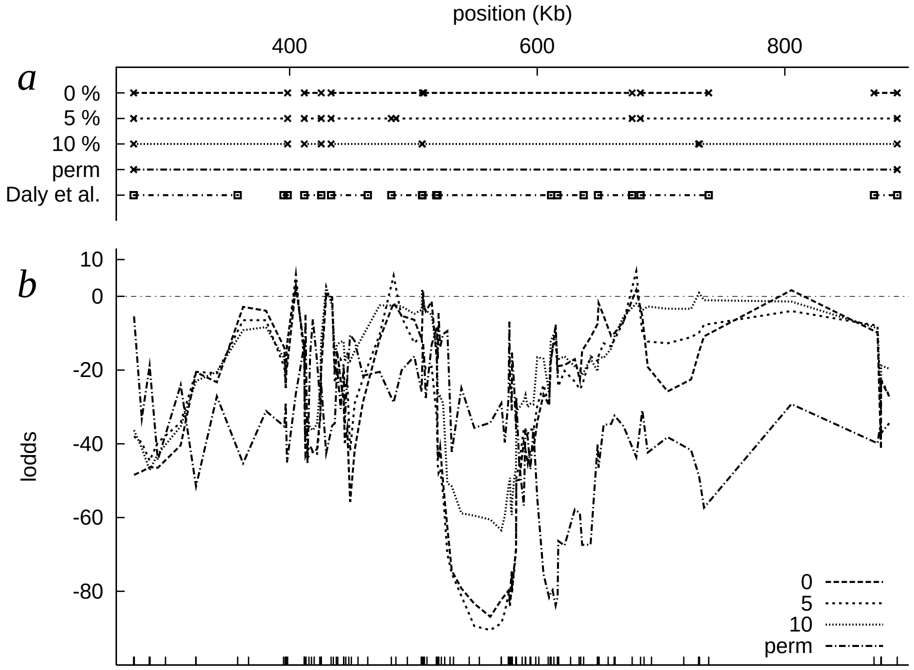

Figure 2: Haplotype block structure in the data of Daly *et al.* when noise is added. The x-axis is the physical location of the marker. (a) The block boundaries reported by Daly *et al.* and the optimal block structures produced by the MDL scoring function when 0, 5, and 10 percent of random noise was added to the data and when the order of markers was randomly permuted. (b) The log odds of the probability of block boundaries for each pair of adjacent markers after adding noise and permuting the order.

settlement sample ($n = 108$ chr.) representing a younger (15–20 generations) population that has gone through a population bottleneck in the 1500s and the third regional subisolate ($n = 108$ chr.), founded by 40 families some 300 years ago, followed by a major expansion.

A total of 45 SNPs were genotyped over 1Mb area on chromosome 1q. Figure 3 shows the results. Five haplotype blocks were identified in three populations. The identified blocks varied in size from 12kb to 361kb. As expected, the haplotype blocks do not differ in different subpopulations of Finland, most probably reflecting the limited set of original founder chromosomes shared by all analyzed populations.

The log odds curve for estimating the probability of boundaries shows that generally the boundaries are stronger for larger values of n; this is also visible in experiments done by sampling from the data of Daly *et al.* (results not

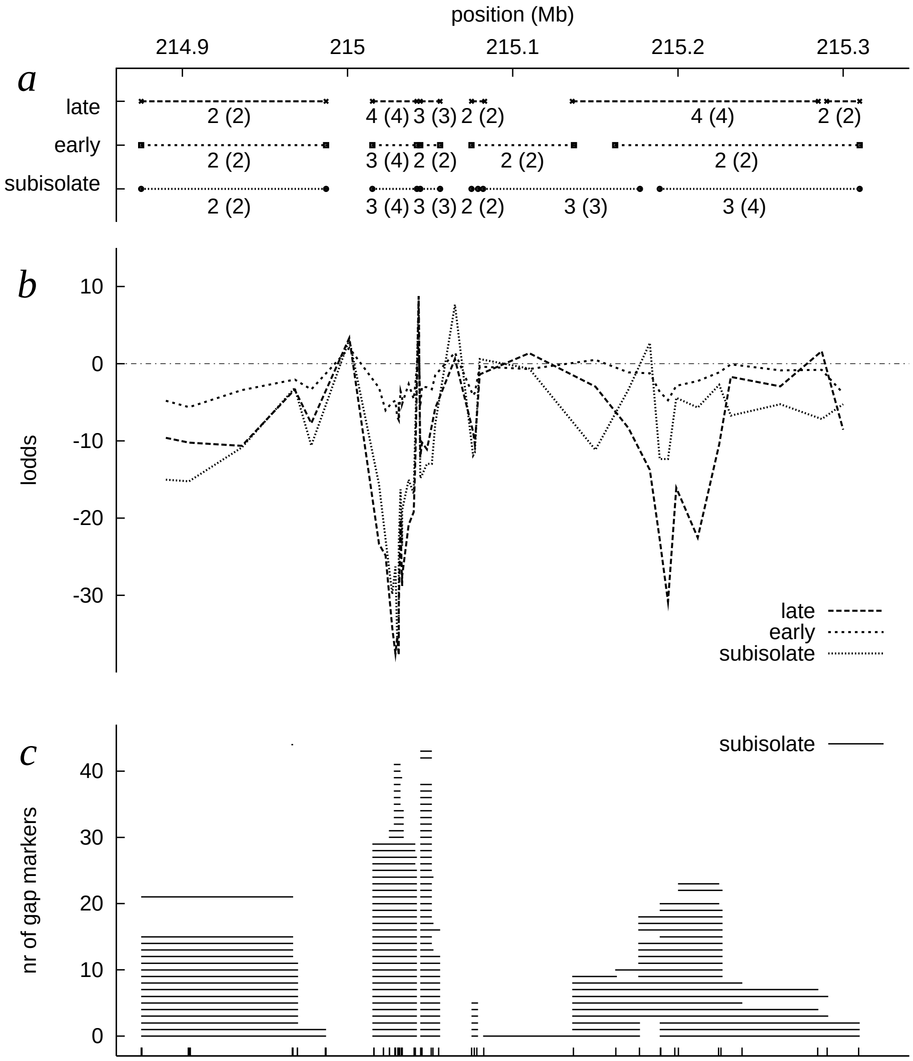

Figure 3: Haplotype block structure in the data from three subpopulations in Finland. Sample sizes: late settlement, $n = 108$; early settlement, $n = 32$; subisolate, $n = 108$. The x-axis is the physical location of the marker. (a) The optimal block structure produced by the MDL scoring function in the three subpopulations. The numbers refer to the number of haplotype classes covering at least 85 percent of haplotypes and, in parenthesis, the full block. (b) The log odds of the probability of block boundaries for each pair of adjacent markers. (c) The subisolate: the optimal segmentation when k markers are allowed to be left outside the blocks, for varying k.

shown).

The significance of block boundaries was also tested using bootstrap methods by investigating whether the optimal segmentation had a block boundary in resampled data sets. The results are similar to the probabilities produced by the probabilistic approach (results not shown).

6 Concluding remarks

We have described a method for defining and finding haplotype blocks based on the use of the minimum description length principle. The clearer the haplotype block structure is, the shorter a description can be given to the data. The coding cost function is such that dynamic programming can be applied to the problem, yielding an $O(np^3)$ algorithm for n observations over p markers. The method can also be used to search for segmentations where a given number of markers are allowed to be left as gaps. We also showed how the MDL principle can be used to obtain probabilities for block boundaries for all pairs of adjacent markers, giving a clear way of evaluating the significance of block boundaries. Experiments on synthetic and real data show that the method produces useful results.

Our method can be extended in many directions. The clustering approach and k-means algorithm could be replaced by closely related but directly probabilistic mixture models and the usual expectation maximization algorithm, respectively. This would also help in modifying the method to handle microsatellite markers. A challenging open problem is to develop an efficient method that is able to discover block structure using genotypic data with unknown phase.

References

1. M.J. Daly, J.D. Rioux, S.F. Schaffner, T.J. Hudson, E.S. Lander, High-resolution haplotype structure in the human genome, *Nature Genetics* **29**, 229–232 (2001).
2. N. Patil, A.J. Berno, D.A. Hinds *et al.*, Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21, *Science* **294**, 1719–1723 (2001).
3. K. Zhang, M. Deng, T. Chen, M.S. Watermanm, F. Sun, A dynamic programming algorithm for haplotype block partitioning, *PNAS* **99**, 7335–7339 (2002).
4. S.B. Gabriel, S.F. Schaffner, H. Nguyen *et al.*, The structure of haplotype blocks in the human genome, *Science* **296**, 2225–2229 (2002).

5. J. Rissanen, Modeling by shortest data description, *Automatica* **14**, 465–471 (1978).
6. J. Rissanen, Stochastic Complexity, *Journal of the Royal Statistical Society* B **49**, 223–239 (1987).
7. M. Li, P. Vitanyi, *An Introduction to Kolmogorov Complexity and its Applications* (Springer-Verlag, New York, 1997).
8. M.H. Hansen, B. Yu, Model Selection and the Principle of Minimum Description Length, *Journal of the American Statistical Association* **96**, 746–774 (2001).
9. W. Li, New stopping criteria for segmenting DNA sequencies, *Physical Review Letters* **86**, 5815–5818 (2001).
10. J.R. Quinlan, R.L. Rivest, Inferring decision trees using the Minimum Description Length principle, *Information and Computation* **80**, 227–248 (1989).
11. P. Kilpeläinen, H. Mannila, E. Ukkonen, MDL learning of unions of simple pattern languages from positive examples, In *Proceedings of the Second European Conference on Computational Learning Theory (EuroCOLT)*, ed. Paul Vitanyi (Springer-Verlag, Berlin, 1995).
12. P. Domingos, The role of Occam's razor in knowledge discovery, *Data Mining and Knowledge Discovery* **3**, 1–19 (1999).
13. R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, 1998).
14. J.S. Liu, C.E. Lawrence, Bayesian inference on biopolymer models, *Bioinformatics* **15**, 38–52 (1999).
15. T. Paunio, J. Ekelund, T. Varilo *et al.*, Genome-wide scan in a nationwide study sample of schizophrenia families in Finland reveals susceptibility loci on chromosomes 2q and 5q, *Human Molecular Genetics* **10**, 3037–3048 (2001).