Lecture notes

Modelling DNA Copy Number Data using HMMs [1]

Susann Stjernqvist, Lund University

1 Introduction to DNA copy number data

The genetic material in human cells is formed in 23 pairs of chromosomes. In each pair, one chromosome is inherited from the mother and the other from the father. The chromosomes consists of DNA, which carries information coded by sequences of pairs of the nucleotide bases A, C, G and T. At each position there are one base pair, either A-T or C-G, and in total we have about $3 \cdot 10^9$ base pairs (bp) in 23 chromosomes. As stated before there are two copies of each chromosome, and thereby also two variants of DNA at each base pair position (except in the sex chromosomes). DNA in cancer cells could however exist in a different number of copies than two. These aberrations usually occur in segments, either short or longer, up to an entire chromosome in length. If the copy number is smaller than two it implies that one or both copies are lost. Similarly if the copy number is larger than two, there are one or several extra copies. By finding the regions with copy number aberrations one can increase the knowledge about the disease, and thereby make better diagnoses and improve treatments. Examples of copy number aberrations are shown in Figure 1- 2.

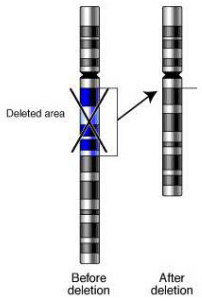

Figure 1: An example of when a short segment of one chromosome is lost.

One technique used to measure the number of copies of DNA is array comparative genomic hybridisation (aCGH). Briefly, the idea of aCGH is to

[1] Since these notes are written independent of the book, the notation deviates somewhat from the rest of the course material.

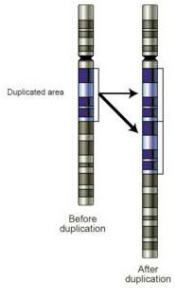

Figure 2: An example of when a short segment of one chromosome is gained.

label many short sequences (denoted probes) of sample DNA and reference DNA with two different fluorescenting dyes, and then measure the intensity ratio between them, when irradiating with a laser. This provides one intensity ratio for each probe. Since the copy number of the reference sample is 2, it follows that when the ratio is 2/2 there are no abberations, i.e. two copies of the sample DNA as well. Further on, the ratio is 1/2 if one copy is lost, 3/2 if one copy is gained and so on. The ratios are translated into $\log_2$-scale putting the normal level at zero. The aCGH method however provides measurement errors to the $\log_2$-ratios, which can be seen in Figure 3 where an example of the data used in Stjernqvist *et al.* (2007) is shown.

2 Discrete-index hidden Markov models

aCGH data has been analysed using several different models and methods, among which various kinds of hidden Markov models are frequently used. Studying Figure 3 it is visible that probes located near each other often tend to have the same copy number, which makes a Markov approach natural. In addition there are a countable number of copy numbers, well suited to be represented by the states of the Markov process. Due to different measurement errors we observe the $\log_2$-ratios in noise, which corresponds to the theory of hidden Markov models.

The model and method described below was originally introduced in Fridlyand *et al.* (2004). Each chromosome is modelled separately and the measurements are ordered in the same order as the probes occur in the chromosomes. Let y_k denote the observed log_2-ratio of probe number k, with $k = 1, \ldots, N$, and N is the total number of probes in that chromosome. In addition let $X = \{X_k\}_{k=1}^{N}$ be a Markov process which corresponds to the hidden copy numbers, with transition probability matrix A, such that

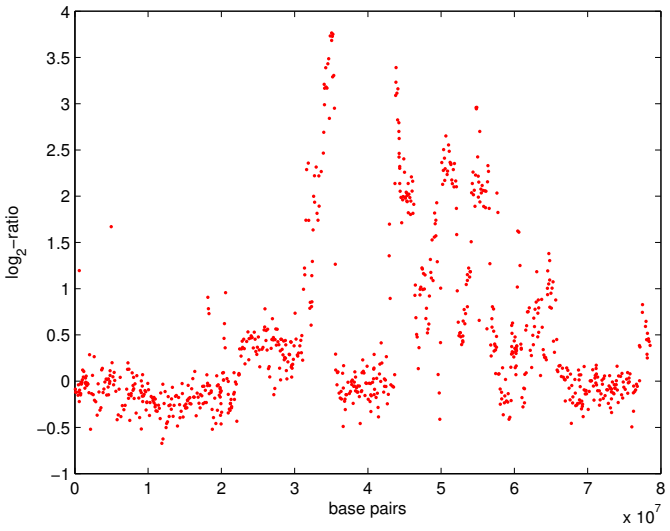

Figure 3: Measured log_2-ratios in chromosome 17 in a cancer cell. Each red dot corresponds to one probe.

$a_{ij} = P(X_k = j|X_{k-1} = i)$. The state space of the Markov chain is $\{1, \ldots, m\}$ and represents the copy numbers of the test sample. The noise is assumed to be Normally distributed such that $Y_k|X_k = i \sim N(\mu_i, \sigma^2)$.

Ideally the mean values μ_i of the states would be equal to $\log_2(i/2)$, where i is an integer, but the data have different sorts of systematic errors causing deviations from this. Instead we choose to estimate the means along with the other parameters. So, in total we wish to estimate the means μ_i, the variance σ^2, the initial probability ρ_i, and the transition probabilities a_{ij}. In addition we would like to reconstruct the Markov process X. For the parameter estimations we use the EM algorithm along with the forward-backward algorithm. Let $\theta = \{\mu_i, \sigma^2, p_{ij}, \rho_i \text{ for } i, j = 1, \ldots, m\}$ and denote by $g_{x_k}(y_k; \theta)$ the conditional density of $Y_k|X_k = x_k$. The complete likelihood is then

$$\begin{aligned} & f_{X_1,\ldots,X_N,Y_1,\ldots,Y_N}(x_1, \ldots, x_N, y_1, \ldots, y_N; \theta) \\ = \; & p_{X_1,\ldots,X_N}(x_1, \ldots, x_N; \theta) f_{Y_1,\ldots,Y_N|X_1,\ldots,X_N}(y_1, \ldots, y_N|x_1, \ldots, x_N; \theta) \\ = \; & \rho_{x_1} \prod_{k=2}^{N} a_{x_{k-1}x_k} \prod_{k=1}^{N} g_{x_k}(y_k; \theta). \end{aligned}$$

Following the theory of the EM algorithm the expectation step is

$$\begin{aligned} Q(\theta,\theta') &= E_\theta[\log f_{X_1,\ldots,X_N,Y_1,\ldots,Y_N}(x_1,\ldots,x_N,y_1,\ldots,y_N;\theta')|y_1,\ldots,y_N] \\ &= \sum_{i=1}^{m} \log \rho'_i P_\theta(X_1=i|y_1,y_2,\ldots,y_N) \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{m} \log a'_{ij} \sum_{k=2}^{N} P_\theta(X_{k-1}=i,X_k=j|y_1,\ldots,y_N) \\ &+ \sum_{i=1}^{m}\sum_{k=1}^{N} \left(-\frac{1}{2}\log(2\pi\sigma'^2) - \frac{(y_k-\mu'_i)^2}{2\sigma'^2}\right) P_\theta(X_k=i|y_1,y_2,\ldots,y_N). \end{aligned}$$

The value of θ' that maximises Q is given by

$$\begin{aligned} \rho'_i &= P_\theta(X_1=i|y_1,\ldots,y_N), \\ p'_{ij} &= \hat{n}_{ij}/\hat{n}_i, \\ \mu'_i &= \frac{\sum_{k=1}^{N} y_k P_\theta(X_k=i|y_1,\ldots,y_N)}{\sum_{k=1}^{N} P_\theta(X_k=i|y_1,\ldots,y_N)}, \\ \sigma'^2 &= \frac{1}{N}\sum_{i=1}^{m}\sum_{k=1}^{N}(y_k-\mu'_i)^2 P_\theta(X_k=i|y_1,\ldots,y_N), \end{aligned}$$

where

$$\begin{aligned} \hat{n}_{ij} &= \sum_{k=2}^{N} P_\theta(X_{k-1}=i,X_k=j|y_1,\ldots,y_N) \\ &= \sum_{k=2}^{N} \frac{\alpha_{k-1}(i)p_{ij}g_j(y_k)\beta_k(j)}{\sum_{s=1}^{m}\sum_{t=1}^{m}\alpha_{k-1}(s)p_{st}g_t(y_k)\beta_k(t)}, \end{aligned}$$

and $\hat{n}_i = \sum_{j=1}^{m} \hat{n}_{ij}$.

Then by including the forward variables $(\alpha_k(i))$ and backward variables $(\beta_k(i))$, the parameters can be estimated. Finally we wish to reconstruct the Markov process. This is performed using the Viterbi algorithm and some results can be found in Figure 4- 5.

3 A continuous-index HMM

As mentioned before there are several other HMMs used to model aCGH data. One example is the method in Stjernqvist *et al.* (2007), which is designed for aCGH data from what is called tiling BAC arrays. Features of such data is that the probes are unevenly spread over the genome, have different lengths and could overlap, i.e one probe starts before the previous probe has ended. The discrete index hidden Markov process described above

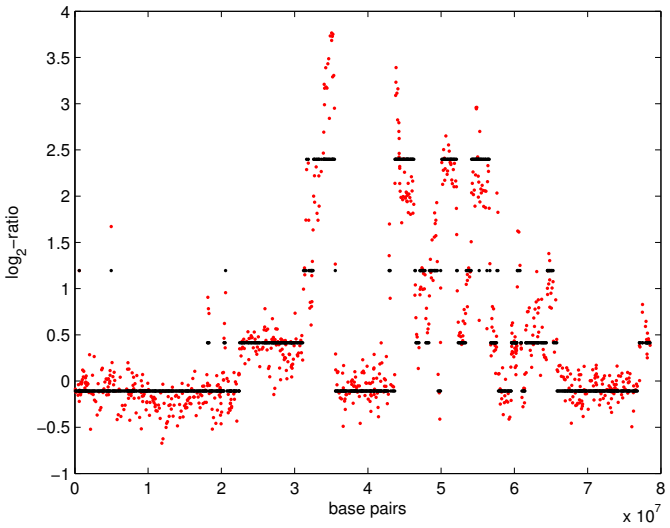

Figure 4: Measured $\log_2$-ratios in chromosome 17 in a cancer cell (red dots) and the Markov process reconstructed using the Viterbi algorithm (black line).

does not include those features so instead a continuous-index hidden Markov process is used. Denote the process $X = (X(t))_{0 \le t \le T}$, where T is the length of the chromosome, and let it take values in a similar state space as the discrete index process, i.e. $\{1, \ldots, m\}$. The dynamics of the Markov process are given by transition rates q_{ij} for $j \neq i$—rather than transition probabilities—where q_{ij} is the rate with which the chain moves from state i to state j; its unit here is bp^{-1}. In addition we let q_i be the total rate out of state i, such that $q_i = \sum_{i \neq j} q_{ij}$. The parameters ρ_i, μ_i, σ^2 are defined similarly as in the discrete index model, as well as y_k, which is the $\log_2$-ratio of probe k. Here however, we note that probe k have a length and is located with start and stop positions denoted t_k^{start} and t_k^{stop} respectively. We also keep the assumption of Normally distributed noise, but includes the length of the probe into the model such that

$$Y_k \sim N\left(\frac{1}{t_k^{stop} - t_k^{start}} \int_{t_k^{start}}^{t_k^{stop}} \mu_{X(t)}\, dt,\, \sigma^2 \right). \tag{1}$$

Due to the overlapping probes it is no longer possible to use the standard EM algorithm and instead an MCEM algorithm is applied. The difference with MCEM is that the E-step is replaced by Monte Carlo simulations of the hidden process. This will not be explained any further here, but those who are interested can read more in Stjernqvist *et al.* (2007). Instead we

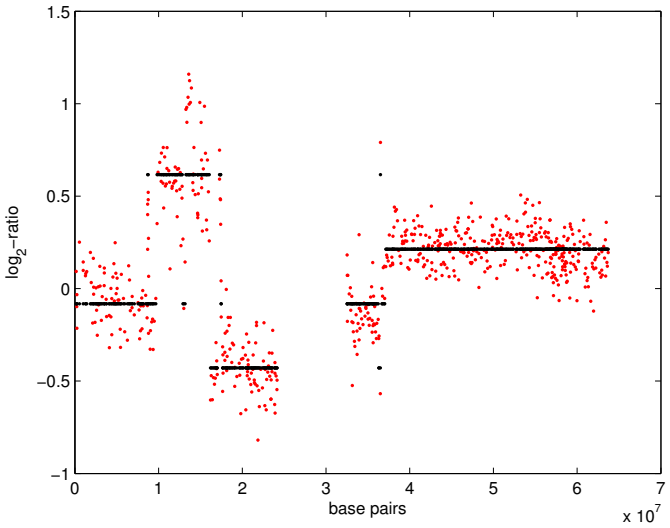

Figure 5: Measured $\log_2$-ratios in chromosome 19 in a cancer cell (red dots) and the Markov process reconstructed using the Viterbi algorithm (black line).

move on to some results, showing one advantage using a probabilistic model like HMM.

It is often interesting to find the positions where the copy number changes state. For example in Figure 6 it is clear that a part of the data is in a different state than the rest. Using the continuous-index model we can estimate the position of the jump. One alternative is to use the Monte Carlo simulations, but we will here focus on a more theoretical alternative. Assume that $[T_1, T_2]$ is an interval such that we know that $X(T_1) = i$, $X(T_2) = j$ and that there is exactly one jump in the interval; assume also that the model parameters are given. Then the conditional density of the jump location, t say, is proportional to

$$q_{ij} \exp(-q_i(t - T_1)) \exp(-q_j(T_2 - t)) \prod_k g(y_k|x) \tag{2}$$

for $T_1 < t < T_2$, and the product is taken over probes k that overlap with the interval $[T_1, T_2]$. This quantity is computed twice using the data in Figure 6 using first $T_1 = 10$ Mbp and $T_2 = 11$ Mbp, and then using $T_1 = 11.5$Mbp and $T_2 = 12.5$ Mbp.

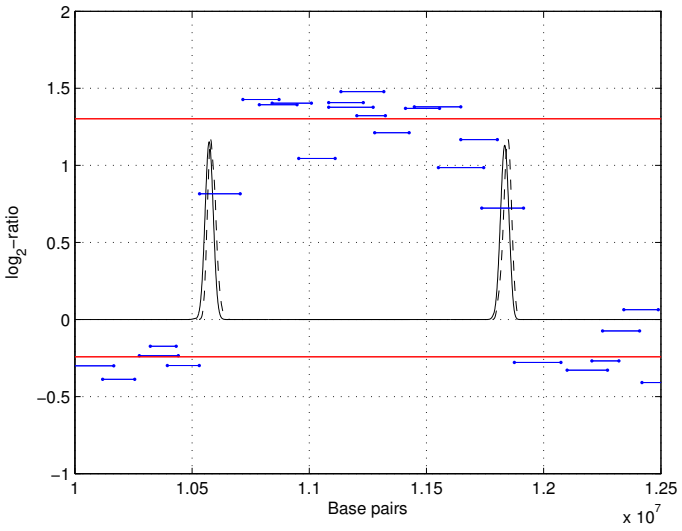

Figure 6: Part of data of chromosome 4 with measured $\log_2$-ratios (short blue lines with dots at the ends) and estimated means of the two states (red lines). The solid lines are jump location estimations from Monte Carlo simulations and the dashed black lines are estimated using Equation 2.

References

Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G. and Jain, A.N. (2004). Hidden Markov models approach to the analysis of array CGH data, *J. Multivar. Anal.*, **90**, 132–153.

Stjernqvist, S., Rydén, T., Sköld, M. and Staaf, J. (2007). Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, **23**, 1006–1014.