Statistical Methods in Physical Mapping

Schbath, Sophie
Institut National de la Recherche Agronomique,
Unité Mathématique, Informatique et Génome,
78026 Versailles, France.
Sophie.Schbath@versailles.inra.fr

Key words: physical mapping, islands, genome coverage, Poisson process.

1 Introduction

A physical mapping project usually aims to produce a physical map of each chromosome of an organism. Such physical map is a set of ordered and overlapping genomic fragments (clones) spanning the entire chromosome. A genomic library of clones is first constructed, generally by partial restriction digest, and it is supposed to represent the chromosome. Then, clones are chosen at random and overlaps are inferred. Overlapping clones are linked into islands or contigs. Several approaches can be used to infer clone overlaps. For instance, clones or clone ends are characterized by a fingerprint (sequence or restriction fragment information) and fingerprints are compared. One declares two clones overlap if their fingerprints are sufficiently similar. For long clones, genomic mapping by anchoring random clones is more adapted. It requires a genomic library of anchors: they are very small genomic sequences assumed to uniquely identify a chromosome location. Thus, clones containing a common anchor overlap.

Statistical questions addressed in a physical mapping project concern the prediction of its progress. How many islands are obtained, what is their mean length, how many clones the islands are composed of, what is the proportion of chromosome covered by the islands ? etc. One is interested in studying the evolution of the physical map with respect to the number of clones studied, to their length, and to some parameters related to the mapping strategy. Lander and Waterman (1988) and Port et al. (1994) have presented such a mathematical analysis for physical mapping by fingerprinting or by end-characterized random clones. In this paper, we restrict ourself to the random anchoring strategy. The main reference is Arratia et al. (1991).

Solving these questions requires to model the clone locations along the genome and the problem falls under the general heading of coverage processes (Hall, 1988). Usually, one models clone locations by a homogeneous Poisson process along the genome with rate $c = LN/G$, where G is the genome length in basepairs, L is the clone length in basepairs and N is the (mean) number of clones. It means that clones are uniformly distributed along the chromosome and c represents the expected number of clones covering a random point (also called the redundancy of coverage). It is possible to relax the fixed length of the clones by considering random lengths independent and identically distributed with a given probability density function. Homogeneity assumptions can also be relaxed by considering a clone length distribution and/or a clone rate c that depend on the location of the clones (see Schbath, 1997 and Schbath *et al.*, 2000). This papers aims to give an idea on the statistical methods used to determine the progress of an STS mapping project. For convenience, we use the homogeneous framework.

In the next section, we set the probabilistic model for the clones and the anchors. Then, we present how to calculate the mean proportion of genome covered by anchored islands. Regarding the length of the paper, we make this choice for two reasons. Biologically, the proportion of genome covered by islands is probably the most important quantity of interest. Mathematically, it gives a nice view on the technics used. This is the purpose of Section 4. Section 3 gives a preliminary result.

2 Probabilistic model and properties

We suppose that clone lengths L are independent and identically distributed random variables with mean length $\mathbb{E}(L)$. We rescale the genome by $\mathbb{E}(L)$ so that the normalized genome corresponds to the interval $(0, g]$ on the real line, with $g = G/\mathbb{E}(L)$. Let f be the probability density function of the normalized clone lengths denoted by $Q = L/\mathbb{E}(L)$. We define the function $\mathcal{F}(u) = \mathbb{P}(Q > u) = \int_u^\infty f(q)dq$. We assume that right-hand ends of clones are distributed independently across the real line $(-\infty, +\infty)$ according to an homogeneous Poisson process with rate $c = N/g$. Recall that N is the mean number of clones used in the mapping project. The anchors are also randomly drawn along the real line according to an homogeneous Poisson process, independent of the clone process, and with rate $a = M/g$; M being the mean number of anchors used.

Recall that a general Poisson process is defined so that the number of points falling into two disjoint sets are independent and moreover, the number of points falling in a given set is Poisson distributed. More precisely for a Poisson process on the real line with rate $\lambda(t)$, the number of points falling into a set A is distributed according to the Poisson distribution with mean $\int_A \lambda(t)dt$.

The main property of an homogeneous Poisson process is that the distances between consecutive points are independent and identically distributed according to an exponential distribution. Let us prove the following related property that will be used later.

Property 1 *Let V_t be the distance forward from a given position t to the next anchor and let W_t be the distance back from t to the previous anchor.*

(i) V_t is exponentially distributed with mean a^{-1}; its probability density function is

defined by

$$g(v) = \begin{cases} a\exp(-av), & v \geq 0, \\ 0, & v < 0. \end{cases}$$

(ii) W_t is exponentially distributed with mean a^{-1}; its probability density function is g.

(iii) V_t and W_t are independent random variables.

Proof. (i) Set $v \geq 0$; V_t is strictly greater than v if and only if no anchors fall into the interval $(t, t+v]$. It occurs with a probability equal to $\exp(-av)$ because the number of anchors falling into $(t, t+v]$ is a Poisson variable with mean av. Therefore, we have

$$\mathbb{P}(V_t \leq v) = 1 - \exp(-av).$$

By differentiating this equality, it gives $g(v) = a\exp(-av)$.

(ii) The proof is similar to the one of (i).

(iii) Set $v \geq 0$ and $w \geq 0$; simultaneously, V_t is strictly greater than v and W_t is strictly greater than w if and only if no anchors fall into $(t, t+v]$ and no anchors fall into $[t-w, t)$. Since these two intervals are disjoint, the anchors falling into $(t, t+v]$ are independent from the ones falling into $[t-w, t)$. Therefore, $\mathbb{P}(V_t > v, W_t > w) = \mathbb{P}(V_t > v)\mathbb{P}(W_t > w)$ and the independence is established.

□

3 Preliminary result

Proposition 1 *The probability $J(x)$ that two points separated by distance x are not covered by a common clone is*

$$J(x) = \exp\left(-c\int_x^{\infty} \mathcal{F}(u)\, du\right).$$

Proof. Set $t \in \mathbb{R}$. A clone of length Q ending at position y covers $[t, t+x]$ if and only if $t + x \leq y$ and $y - Q \leq t$. Averaging over the distribution of Q gives that the probability that a clone ending at y covers $[t, t+x]$ is $\mathcal{F}(y-t)\mathbb{I}\{y \geq t+x\}$. Since the right-hand ends of clones is a Poisson process with rate c, the process of right-hand ends of clones covering $[t, t+x]$ is a Poisson process with inhomogeneous rate $c\,\mathcal{F}(y-t)\mathbb{I}\{y \geq t+x\}$ (Marking Theorem, Kingman, 1983). The total number of clones covering $[t, t+x]$ is then a Poisson variable with mean $\int_{\mathbb{R}} c\,\mathcal{F}(y-t)\mathbb{I}\{y \geq t+x\}\, dy$ and $J(x)$ is the probability that this Poisson variable is equal to zero :

$$J(x) = \exp\left(-c\int_{t+x}^{+\infty} \mathcal{F}(y-t)\, dy\right).$$

The change of variable $u = y - t$ leads to the result.

$\square$

Remark 1 *The probability that a given location is covered by a clone is given by* $1 - J(0) = 1 - \exp(-c)$.

4 Proportion of genome covered by anchored islands

Because of the homogeneity assumption, the mean proportion of chromosome not covered by anchored islands is also the probability that a given location t on the chromosome is not covered by an anchored island. To calculate this probability denoted by $r_0(t) = r_0$, we start working conditionally with V_t the distance forward from t to the next anchor and with W_t the distance back from t to the previous anchor. Given V_t and W_t, the location t is not covered by an anchored island if and only if no clones cover simultaneously the locations $t - W_t$ and t, and no clones cover simultaneously the locations t and $t + V_t$. Because clones covering $[t - W_t, t]$ are not independent from the ones covering $[t, t + V_t]$, we write the previous conditions as follows. The location t is not covered by an anchored island if and only if no clones cover simultaneously t and $t + V_t$ (event E_1) and no clones ending in $[t, t + V_t)$ cover $t - W_t$ (event E_2). Now the events E_1 and E_2 are independent because E_1 is related to clones ending in $[t + V_t, +\infty)$, E_2 is related to clones ending in $[t, t + V_t)$ and the right-hand ends of clones is a Poisson process. We then have

$$\mathbb{P}(t \text{ is not covered by an anchored island} \mid V_t, W_t) = \mathbb{P}(E_1 \mid V_t, W_t)\mathbb{P}(E_2 \mid V_t, W_t).$$

Proposition 1 gives $\mathbb{P}(E_1 \mid V_t, W_t) = J(V_t)$. To calculate the conditional probability of E_2, we consider a clone ending at position y. Given its length Q, this clone ends in $[t, t + V_t)$ and covers $t - W_t$ if and only if $t \le y < t + V_t$ and $y - Q < t - W_t$. Therefore, a clone ending at y ends in $[t, t + V_t)$ and covers $t - W_t$ with probability

$$p(y) = \mathcal{F}(y - t + W_t)\mathbb{1}\{t \le y < t + V_t\}.$$

Because the right-hand ends of clones is a Poisson process with rate c, the Marking Theorem ensures that the right-hand end of clones ending in $[t, t + V_t)$ and covering $t - W_t$ is an inhomogeneous Poisson process with rate $cp(\cdot)$. The total number of such clones is then a Poisson variable with mean $\int_{\mathbb{R}} cp(y)\, dy$. It follows

$$\begin{aligned} \mathbb{P}(E_2 \mid V_t, W_t) &= \exp\left(-\int_t^{t+V_t} c\,\mathcal{F}(y - t + W_t)\, dy\right) \\ &= \exp\left(-c\int_{W_t}^{V_t + W_t} \mathcal{F}(u)\, du\right). \end{aligned}$$

Using Proposition 1, we can write $\mathbb{P}(E_2 \mid V_t, W_t) = J(W_t)/J(V_t + W_t)$ and

$$\mathbb{P}(t \text{ is not covered by an anchored island} \mid V_t, W_t) = \frac{J(W_t)J(V_t)}{J(V_t + W_t)}.$$

We then have to average the previous conditional probability according to the distribution of the couple (V_t, W_t). Using Property 1 leads to

$$r_0 = r_0(t) = \int_{v=0}^{\infty} \int_{w=0}^{\infty} \frac{J(w)J(v)}{J(v+w)} a^2 \exp -a(v+w) dv\, dw.$$

REFERENCES

ARRATIA, R., LANDER, E. S., TAVARÉ, S. and WATERMAN, M. S. (1991). Genomic mapping by anchoring random clones: A mathematical analysis. *Genomics* **11** 806–827.

HALL, P. (1988). *Introduction to the Theory of Coverage Processes.* Wiley, New York.

KINGMAN, J. F. C. (1993). *Poisson processes.* Oxford-University Press.

LANDER, E. S. and WATERMAN, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2** 231–239.

PORT, E., SUN, F., MARTIN, D. and WATERMAN, M. S. (1995). Genomic mapping by end-characterized random clones: A mathematical analysis. *Genomics* **26** 84–100.

SCHBATH, E. (1997). coverage processes in physical mapping by anchoring random clones *Journal of Computational Biology* **4** 61–82.

SCHBATH, E., BOSSARD, N. and TAVARÉ, S. (2000). The effect of nonhomogeneous clone length distribution on the progress of an STS mapping project. *Journal of Computational Biology* **7** 47–57.