Coverage Processes in Physical Mapping by Anchoring Random Clones

Sophie Schbath
INRA

February 10, 1997

[summary by Mireille Régnier]

1. Introduction

A complete physical map of the DNA of an organism consists of ordered overlapping fragments spanning the entire genome. A large number of fragments, called *clones*, are chosen at random from a library in which the entire genome is represented. The anchoring approach, efficient for large clones and then for very large genomes, uses an additional random genomic library of *anchors* and is based on the anchor-content of the clones. These anchors consist of any short sequence of DNA that occurs exactly once in the genome. Thus, clones containing an anchor in common overlap. Clones that overlap are then assembled into *islands* which cover some regions of the genome.

To plan a physical mapping project, it is therefore important to study, for instance, the proportion of the genome covered by islands, the number and the length of islands, with respect to the number of clones and anchors considered. To study the statistical properties of the islands, we must model the clones and anchors processes. This problem is typically part of the theory of coverage processes [6]. The aim of this paper is to provide a mathematical analysis of physical mapping by anchoring random clones, in a general model.

2. Notation and preliminary results

We consider the genome linearly as a discrete sequence of bases of length G. We assume that clones have independent and identically distributed lengths L, with mean $\mathcal{E}L$. We define $g = G/\mathcal{E}L$. Since the anchor length is very small compared to the clone length, the anchors will be considered to be points. We assume that right ends of clones occur on the real line according to a non-homogeneous Poisson process of rate $\alpha(t)$, and we label them $\{C_i, i \in \mathbb{Z}\}$ such that $\cdots < C_{-1} < C_0 \leq 0 < C_1 < \cdots < C_{N(g)} \leq g < C_{N(g)+1} < \cdots$; $N(t) \equiv N((0,t])$ denotes the counting process of clones and represents the number of clones ending in $(0,t]$. Similarly, the anchors occur according to a non-homogeneous Poisson process of rate $\lambda(t)$ and are labelled $\{A_j, j \in \mathbb{Z}\}$ such that $\cdots < A_{-1} < A_0 \leq 0 < A_1 < \cdots < A_{M(g)} \leq g < A_{M(g)+1} < \cdots$; $M(t) \equiv M((0,t])$ denotes the counting process of anchors and represents the number of anchors in $(0,t]$. We assume the process of anchors and the process of clones are independent, and the rates α and λ are positive functions such that their integrals are finite on bounded sets, but are not finite on unbounded sets.

Proposition 1. *The probability $J(t;x)$ that the points t and $t+x$ $(x>0)$ are not covered by a common clone is*

$$J(t;x) = \exp\left(-\int_x^\infty \alpha(t+u)\mathcal{F}(u)du\right).$$

3. Main results

We introduce the following terminology: a clone containing an anchor is an anchored clone, and the clones containing one or more common anchors are assembled into anchored islands An anchored island can be composed of a unique anchored clone. A region between two anchored islands is an ocean.

Several theorems give properties of unanchored clones, anchored islands and notably the singleton anchored islands, composed of a unique anchored clone. The next theorem is about the proportion of oceans, in other words the proportion of the genome not covered by anchored islands.

Theorem 1. *The probability $r_0(t)$ that t is not covered by any anchored island is*

$$r_0(t) = \int_0^\infty \int_0^\infty \frac{J(t;v)J(t-w;w)}{J(t-w;v+w)} \lambda(t+v)\lambda(t-w)\exp\left(-\int_{t-w}^{t+v}\lambda(x)dx\right) dv\, dw.$$

The mean proportion r_0 of the genome not covered by anchored islands is

$$r_0 = \frac{1}{g}\int_0^g r_0(t)\, dt.$$

Under probabilistic assumptions that are satisfied for periodic rates α and λ, the author states the weak law of large numbers for the number of anchored islands and the number of anchored clones. Moreover, the length of the anchored islands, the mean number of clones and the mean number of anchors in an anchored island are derived. Notably, we get:

Theorem 2. *(i) The process of anchors covered by clones is a Poisson process with rate given by*

$$\nu(t) = \lambda(t)(1 - J(t;0));$$

(ii) the average number of anchors per anchored island ending in $(0,g]$, denoted by $\overline{H}_g$, is such that

$$\overline{H}_g - \frac{\int_0^g \lambda(t)(1-J(t;0))\, dt}{\int_0^g \alpha(t)p_1(t)\, dt} \overset{\Pr}{\to} 0 \quad as \quad g \to \infty.$$

where $p_1(t)$ is the probability that a clone ending at t is the rightmost clone of an anchored island.

Whereas an ocean is a region not covered by anchored islands, an actual ocean is defined to be a region not covered by clones.

Theorem 3. *The probability $p_3(t;x)$ that an anchored island ending at t is followed by an actual ocean of length at least x is*

$$p_3(t;x) = J(t+x;0)\exp\left(-\int_t^{t+x}\alpha(u)du\right)\frac{1-q_1(t)}{p_1(t)}.$$

where $q_1(t)$ is the probability that a clone ending at t contains no anchor.

4. Applications

The author presents numerical results for a genome of length $G = 100,000$ kb, 2300 clones of fixed length $L = 250$ kb and 500 anchors, that corresponds approximately to the physical mapping project of *Arabidopsis thaliana* genome [5]. The normalized genome length is $g = 400$, and the mean number of clones and anchors in $(0,g]$ are respectively $\mathcal{E}N \equiv \mathcal{E}N(g) = 2300$ and $\mathcal{E}M \equiv \mathcal{E}M(g) = 500$.

We focus on four quantities, namely the mean number of unanchored clones, the mean number of anchored islands, the average length of anchored islands and the mean proportion of the genome not covered by anchored islands. The first non-homogeneous model assumes each rate is piecewise constant and can take two values. It means the genome is composed of alternating rich and poor regions of clones, and of anchors. This is the *hotspot* model introduced by [4]. The second model allows some *sinusoidal fluctuations* around the constant rates.

4.1. **Hotspot model.** These are the main trends. When the hotspots are in phase, the number of unanchored clones decreases as the level of clone hotspot increases. As both of the hotspot levels increase, the mean number of anchored islands and the mean length of an anchored island decrease, leading to an increase of the mean proportion of the genome not covered by anchored islands.

When the hotspots are completely out of phase, the number of unanchored clones increases as both of the hotspot levels increase. As the level of clone hotspot increases, the mean number of anchored islands increases and then drops, whereas it decreases as the level of anchor hotspot increases. The mean length of anchored islands decreases as the level of clone hotspot increases. The mean proportion of the genome not covered by anchored islands increases as the levels of hotspot increase, rather dramatically when the level of clone hotspot increases.

The most important effect is undeniably that the more the model departs from the stationary model, the greater the mean proportion of genome not covered by anchored islands. Finally, the simulation results given by [4] and the theoretical calculations are quite close.

4.2. **Sinusoidal model.** We consider the following rates

$$\alpha(t) = \alpha_0 + \alpha_1 \sin(\alpha_2 t), \qquad \lambda(t) = \lambda_0 + \lambda_1 \sin(\lambda_2 t).$$

The mean number of unanchored clones seems not to depend on α_1 and increases as λ_1 increases. Increasing either α_1 or λ_1 has the same effect on the average length of anchored islands and the mean proportion of genome not covered by anchored islands: the length decreases whereas the proportion increases. The stationary case, $\alpha_1 = \lambda_1 = 0$ minimizes the mean number of anchored islands and the mean proportion of genome not covered by anchored islands and maximizes the average length of anchored islands.

5. Conclusion

Undeniably, the inhomogeneity in the clone and anchor locations along the genome substantially changes the predictions in a physical mapping project. Since the goal of a physical mapping project (at least the first step) is to obtain few long anchored islands and a small proportion of genome not covered by anchored islands, the previous applications clearly show that using homogeneous Poisson processes for clones and anchors provides an overly optimistic assessment of the progress of the mapping project. The difficulty in practice remains to model the inhomogeneity occurring in the genome. The detection of regions rich or poor in restriction sites involved in the cloning process, for instance, could be a first step in characterizing some hotspots. Since longer DNA sequences are becoming available, modelling heterogeneity has become an important problem in sequence analysis.

References

[1] Arratia (R.), Lander (E. S.), Tavaré (S.), and Waterman (M. S.). – Genomic mapping by anchoring random clones: A mathematical analysis. *Genomics*, vol. 11, 1991, pp. 806–827.

[2] Barillot (E.), Dausset (J.), and Cohen (D.). – Theoretical analysis of a physical mapping strategy using random single-copy landmarks. *Proceedings of the National Academy of Sciences of the USA*, vol. 88, 1991, pp. 3917–3921.

[3] Chumakov (I.), Rigault (P.), Guillou (S.), Ougen (P.), Billaut (A.), Guasconi (G.), Gervy (P.), Le Gall (I.), Soularue (P.), and Grinas (L.). – Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature*, n° 359, 1992, pp. 380–387.

[4] Ewens (W. J.). – *Simulation results for anchored clones.* – Research Report n° 252, Department of Mathematics, Monash University, Australia, 1996. 7 pages.

[5] Ewens (W. J.), Bell (C. J.), Donnelly (P. J.), Dunn (P.), Matallana (E.), and Ecker (J. R.). – Genome mapping with anchored clones: Theoretical aspects. *Genomics*, vol. 11, 1991, pp. 799–805.

[6] Hall (Peter). – *Introduction to the Theory of Coverage Processes.* – John Wiley & Sons Inc., New York, 1988, *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*, xx+408p.

[7] Hudson (T.), Stein (L.), Gerety, Ma (J.), Castle (A.), Silva (J.), Slonim (D.), Baptista (R.), Kruglyak (L.), Xu (S.), Hu (X.), Colbert (A.), Rosenberg (C.), Reeve-Daly (M.), Rozen (S.), Hui (L.), Wu (X.), Vestergaard (C.), Wilson (K.), Bae (J.), Maitra (S.), Ganiatsas (S.), Evans (C.), DeAngelis (M.), Ingalls (K.), Nahf (R.), Horton (L.), Oskin Anderson (M.), Collymore (A.), Ye (W.), Kouyoumjian (V.), Zemsteva (I.), Tam (J.), Devine (R.), Courtney (D.), Renaud (M.), Nguyen (H.), O'Connor (T.), Fizames (C.), Fauré (S.), Gyapay (G.), Dib (C.), Morissette (J.), Orlin (J.), Birren (B.), Goodman (N.), Weissenbach (J.), Hawkins (T.), Foote (S.), Page (D.), and Lander (E. S.). – An STS-based map of the human genome. *Science*, vol. 270, 1995, pp. 1945–1954.

[8] Karlin (S.) and Macken (C.). – Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *Journal of the American Statistical Association*, vol. 86, 1991, pp. 27–35.

[9] Lander (E. S.) and Waterman (M. S.). – Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, vol. 2, 1988, pp. 231–239.

[10] Lee (W.). – *A mathematical analysis of genome physical mapping.* – Master's thesis, Department of Mathematics, Unversity of Southern California, 1992.

[11] Marr (G. T.), Yan (X.), and Yu (Q.). – Genomic mapping by single copy landmark detection: A predictive model with a discrete mathematical approach. *Mamm. Genome*, vol. 3, 1992, pp. 644–649.

[12] Nelson (D. O.) and Speed (T. P.). – Predicting progress in directed mapping projects. *Genomics*, vol. 24, 1994, pp. 41–52.

[13] Olson (M. V.), Dutchik (J. E), Graham (M. Y.), Brodeur (G. M.), Helms (C.), Frank (M.), MacCollin (M.), Scheinman (R.), and Frank (T.). – Random-clone strategy for genomic restriction mapping in yeast. *Proceedings of the National Academy of Sciences of the USA*, vol. 83, 1986, pp. 7826–7830.

[14] Port (E.), Sun (F.), Martin (D.), and Waterman (M. S.). – Genomic mapping by end-characterized random clones: A mathematical analysis. *Genomics*, vol. 26, 1995, pp. 84–100.

[15] Torney (D. C.). – Mapping using unique sequences. *Journal of Molecular Biology*, vol. 217, 1991, pp. 259–264.

[16] Zhang (M. Q.) and Marr (T. G.). – Genome mapping by nonrandom anchoring: A discrete theoretical analysis. *Proceedings of the National Academy of Sciences of the USA*, vol. 90, 1991, pp. 600–604.