Adventures in Modeling Cancer Evolution

Prof. Simon Tavaré
Irving Institute for Cancer Dynamics
Columbia University

Typed by Yuelin Li

2019 APMA colloquium

1 Introduction

Irving Institute for Cancer Dynamics focuses on quantitative and tech dev aspects of cancer research. Our discussions ongoing with Mathematics Sciences, Biological Sciences, Computer Science, Biomedical Engineering...
What is cancer genomics about? A disease of the genome. Sampling a tumor: sample multiple cancer glands from left/right tumour side, then isolated. How are the data used? Cancer data used in different ways to Smaller numbers of replicates Study cancer at several levels.
- (Cellular) identify biomarkers, predict evolution choose treatments, assess response
- (Patient) which combination therapy, given tumor, genes, other patient data ("precision medicine") early detection
- (Population) effects of lifestyle, geography, cancer type

2 Cancer Evolution

Tumour heterogeneity is about all differences between tumour either within populations or within individuals or within a tumour, and is metastasized. In this talk we focus on the path about the genetic variation within a given tumour. Mathematical modeling in A vibrant field of applied mathematics:

Bellomo, Li, Maini(2008) On the foundations of cancer modeling: selected topics, speculations and perspectives.

Cristini, Lowengrub (2010) Multiscale modeling of cancer.

Chaplain MAJ. Powathil(2015) Multiscale modelling of intracellular heterogeneities in chemotherapy treatment.

Focus in the talk: cancer genomics and cancer evolution.

Main ingredients: inference for stochastic processes

Population genetics principles:

Study the effects of mutation, selection, variation, recombination on the structure of genetic variation in natural populations.

Allow for demographic effects such as migration, admixture, subdivision and fluctuations in population size.

With the advent of molecular data in the early 1970s, paradigm changed from prospective to retrospective.

Example:

Two data sets involving fruit flies: Drosophila allozyme frequency data

D. tropicalis Esterase-2 locus [$n = 298$]

$$234, 52, 4, 4, 2, 1, 1$$

D. simulans Esterase-C locus [$n = 308$]

$$91, 76, 70, 57, 12, 1, 1$$

What is expected under neutrality? Sewall Wright argued that the observations do not agree at all with the equal frequencies expected for neutral alleles in enormously large population. What was needed: the null distribution of the numbers C(n) = (C1(n), ... , Cn(n)) of alleles represented j times in a sample of size n.

The Ewens Sampling Formula [1972]:

The distribution of the counts in a sample of size n with mutation parameter θ :

$$\begin{aligned} &\mathbb{P}\left[C_j(n) = c_j, j = 1, 2, \ldots, n\right] \\ &= \quad \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j!} \end{aligned}$$

if $c_1 + 2c_2 + \cdots + nc_n = n$.

Consequences:

Number of types, $K = C_1(n) + \ldots + C(n)$ ic sufficient for θ. Therefore can use conditional distribution of $\boldsymbol{C}$ given K for testing adequacy of model. Watterson (1977,1978) suggested

using the homozygosity statistic $F = \sum_{j=1}^{n}(j/n)^2 C_j(n)$ as a test statistic for neutrality. Its distribution can be simulated very rapidly using Chinese Restaurant Process or Feller Coupling. and a rejection method with $\theta = K/\log(n)$.
D. tropicalis Esterase-2 locus [$n = 298$]

$$234, 52, 4, 4, 2, 1, 1 \quad F = 0.647, 95\%(0.226, 0.818)$$

D. simulans Esterase-C locus [$n = 308$]

$$91, 76, 70, 57, 12, 1, 1 \quad F = 0.236.95\%(0.244, 0.824)$$

Beginning of statistical theory for estimators of population quantities like θ: geneticists were using that part of tho data least informative for θ. The Maximum Likelihood Estimator $\hat{\theta}_n$ of θ is asymptotically Normal with mean θ variance proportional to $1/\log n$. The slow rate is because of dependence.
This will lead to ...
- Measure-valued processes lambda
- Inference for stochastic processes including ABC (Approximate Bayesian Computation)
- Coalescent (and related branching process) simulators
- Applications to statistical genetics, human disease cancer
Non-parametric Bayesian inference clustering - Useful paradigm for cancer, in which the individuals are cells
Clonal progression...
Another look at sequence data

$$\begin{array}{c|ccccccc} h_1 & 0 & 1 & 1 & 0 & 0 & \cdots & 0 \\ h_2 & 0 & 1 & 1 & 0 & 0 & \cdots & 0 \\ h_3 & 0 & 1 & 1 & 0 & 1 & \cdots & 1 \\ h_4 & 1 & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ h_n & 1 & 0 & 0 & 1 & 0 & \cdots & 0 \end{array}$$

where 0 denotes ancestral type, 1 the mutant type. The rows are haplotypes, the columns are sites of mutations. f_i = number of mutant copies at ith site (known), K = number of distinct haplotypes (not known). We compute
M_b = the number of sites carrying b copies of the mutant
$= \#\{l|f_l = b\}, b = 1, 2, ..., n-1$, $M(n) = (M_1, M_2, ..., M_{n-1})$ is observed SFS.

Some explicit results are known for features of coalescent models, including constant and varying population size. For example, in the constant population size setting, the number $C_j(n)$ of haplotypes appearing j times in the sample has the ESF with parameter θ. We call $C(n)$ the haplotype frequency spectrum (HFS). For many settings, including most branching processes, simulation is required.

For the SFS, there are corresponding results, the simplest of which is

$$\mathrm{E}M_b = \frac{\theta}{b}, b = 1, 2, \ldots, n-1$$

and many papers have discussed other features of the counts. For example, Dahmer and Kersting (2015) showed that

$$M(n) \Rightarrow (P_1, P_2, \ldots) \qquad \text{as } n \to \infty$$

where the P_i are independent Poisson random variables with $\mathbb{E}(P_i) = \theta/i$.

Inference for the coalescent

- Mutation rate θ often estimated from number of segregating sites, S.
- Joint distribution of (K, S) has been studied (Griffiths, 1982)
- Inference methods, including MLEs, from Griffiths, Felsenstein, ... (1990s)
- Inference often based on summary statistics.

Inference based on the SFS summary statistics arise in several settings. One is inference for Pool-Seq data. This is a whole-genome method for sequencing pools of individuals, and provides a cost-effective alternative to sequencing individuals separately. From such data, we can compute the SFS but we do not know haplotypes.

Problem: infer distribution of K, θ from the SFS. (in a Bayesian setting) (is there another way?)

To compute $\mathcal{L}(K, \theta \mid M(n))$ We will focus on a slightly different version, motivated by of mutations S is large. Thus we scale and bin the SFS by computing the fractions p_i of sites which have relative frequency of the mutant in ranges determined by bins $B_1, B_2, \ldots B_r$. We are then after

$$\mathcal{L}(K, \theta \mid (p_1, \ldots, p_r))$$

The combinatorics are not simple here and explicit results seem hard to come by.

Approximate Bayesian Computation

- Appeared in population genetics literature some 20 years ago
- Designed for problems in which likelihoods are hard (or impossible) to compute. but stochastic mode easy to simulate
- Simplest version is based on rejection method:

Simulate θ from prior π; Simulate data $\mathcal{D}'$ from process with parameter θ; Accept θ as draw from $f(\theta \mid \mathcal{D})$ if $\mathcal{D}' = \mathcal{D}$.

then repeat.

The problem is it never works! The last step never hits the target.

Accept θ if $\rho(\mathcal{D}', \mathcal{D}) < \epsilon \to$ Choice of metric ρ and ϵ crucial.

Value of ϵ measures trade off between comoutability and accuracy

Generates observations from $f(\theta \mid \rho(\mathcal{D}' \cdot \mathcal{D}) < \epsilon$

In practice, compare summary statistics of the data with those of the simulations. Theoretical problem: methods for identifying approximately sufficient statistics; Practical problem: given a set of statistics, how do we choose good ones?

Now many versions of these schemes, including MCMC without likelihoods and sequential Monte Carlo.

Some challenges in cancer sequencing.

Pooled samples of cells of mixed type are sequenced - a mixture of tumor and non-tumor

cells - experiments are rather like pool-seq, but the size n is not known.
Converting read depths from sequencing to site frequencies is difficult.
We want to know the number of clones in the sample (akin to K), and if there are signatures of selection.

Inference from SFS
We need to model joint prior for (n, θ)
Priors for $\theta \sim U(200, 300), n \sim U(100.300)$, Bins are $(0.0, 0.1], (0.1, 0.2], \ldots, (0.9, 1.0)$ We want $\mathcal{L}(K, n, \theta \mid (p_1, \ldots, p_r))$
- Generated 500,000 samples

$$\rho\left(p_i, p_i^{\text{obs}}\right) = \frac{1}{2} \sum_{i=1}^{10} \left| p_i - p_i^{\text{obs}} \right|$$

- The 1% point of ρ values is 0.46.