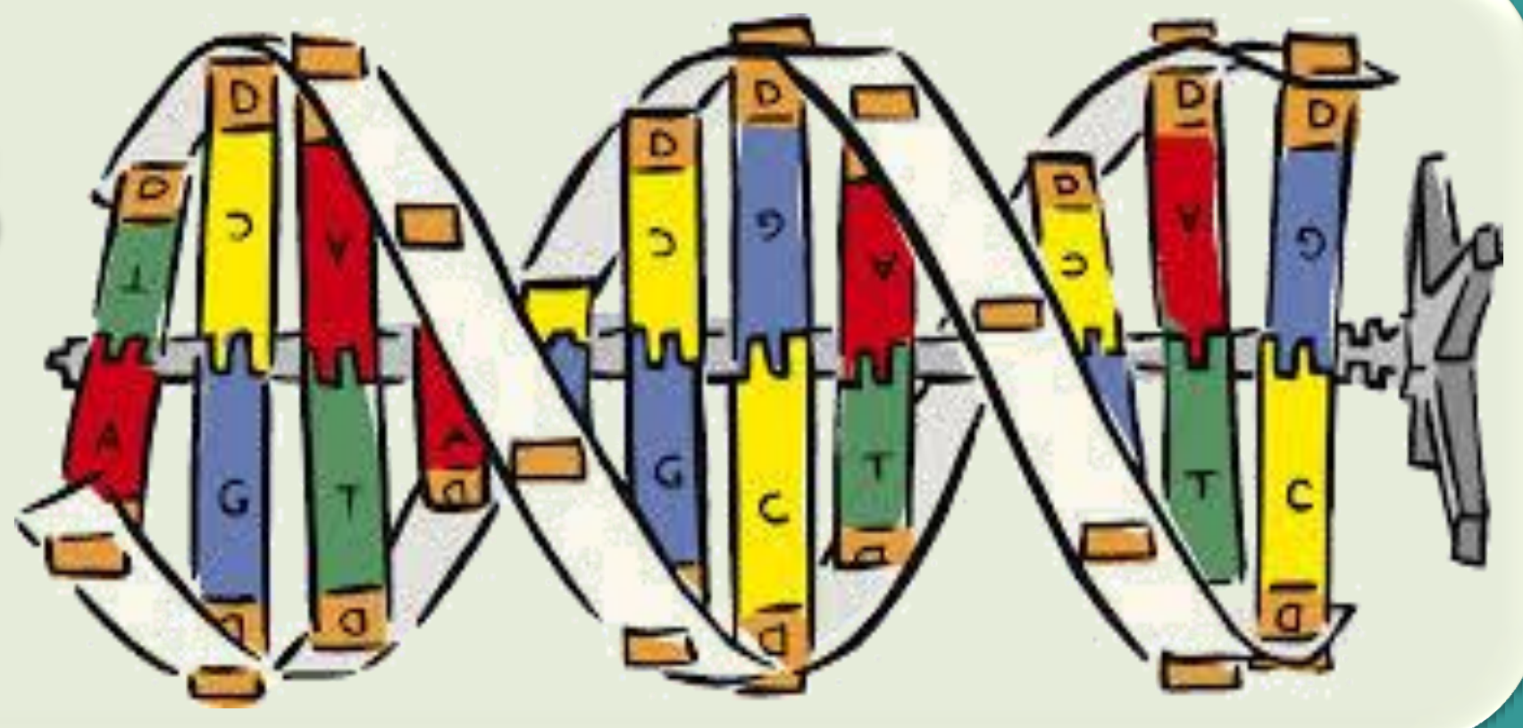
Statistical Models for Discrete Data of DNA samples

Sanjay Man Tamrakar, University of Texas at Tyler
Mentor: Dr. Nathan Smith



Background

Deoxyribonucleic acid, or DNA, is the hereditary material in all living beings, which stores information as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The order, or sequence, of these bases determines the information available for building and maintaining an organism. DNA bases pair up with each other, A with T and C with G, to form units called base pairs. DNA has a double-helical structure, somewhat like a ladder, with the base pairs forming the ladder's rungs. We can use certain statistical models of probability distributions to generate random sequences for discrete data of DNA samples.

Introduction

Let us assume that a process produces a DNA sequence of length N = 49.
CTCACGTGATGAGAGCATTCTCAGACCGTGACGCGTGTAGCAGCGGCTC

The process uses three tetrahedral dice. Two of the dice are loaded and one is fair. The probabilities of rolling the four letters are known to us which are as follows:

	A	C	G	T
First die (θ_1)	0.15	0.33	0.36	0.16
Second die (θ_2)	0.27	0.24	0.23	0.26
Third die ($1-\theta_1-\theta_2$)	0.25	0.25	0.25	0.25

Let p_A, p_C, p_G, and p_T denote the probabilities that the process will generate any of the four letters.

$$p_A = -0.10\,\theta_1 + 0.02\,\theta_2 + 0.25 \qquad p_C = 0.08\,\theta_1 - 0.01\,\theta_2 + 0.25$$
$$p_G = 0.11\,\theta_1 - 0.02\,\theta_2 + 0.25 \qquad p_T = -0.99\,\theta_1 + 0.01\,\theta_2 + 0.25$$

Then the probability of the whole sequence is given by:-

$$p(\text{seq.}) = p_C p_T p_C p_A p_C p_G \ldots\ldots p_C$$
$$= p_A^{10} p_C^{14} p_G^{15} p_T^{10}$$

Maximum Likelihood

R.A. Fisher developed the theory of maximum likelihood estimator. The likelihood function is the probability of the sample, considered as a function of an unknown parameter θ. If we have a large sample, the parameter value maximizing this function will often yield an excellent estimator of θ.

Let $X_1, X_2, X_3, \ldots\ldots X_n$ have joint density denoted by

$$f_\theta(x_1, x_2, \ldots\ldots\ldots x_n) = f(x_1, x_2, \ldots\ldots, x_n|\theta)$$

If we are given observed value $X_1 = x_1$, $X_2 = x_2$, ………., $X_n = x_n$, then the likelihood of θ is the function

$lik(\theta) = f(x_1, x_2, \ldots\ldots, x_n|\theta)$, considered as a function of θ.

The maximum likelihood estimate is the estimator of θ that maximizes the likelihood function of θ. The log-likelihood function is obtained by taking the logarithm on the likelihood function, which is also used because the logarithm is a monotonically increasing function.

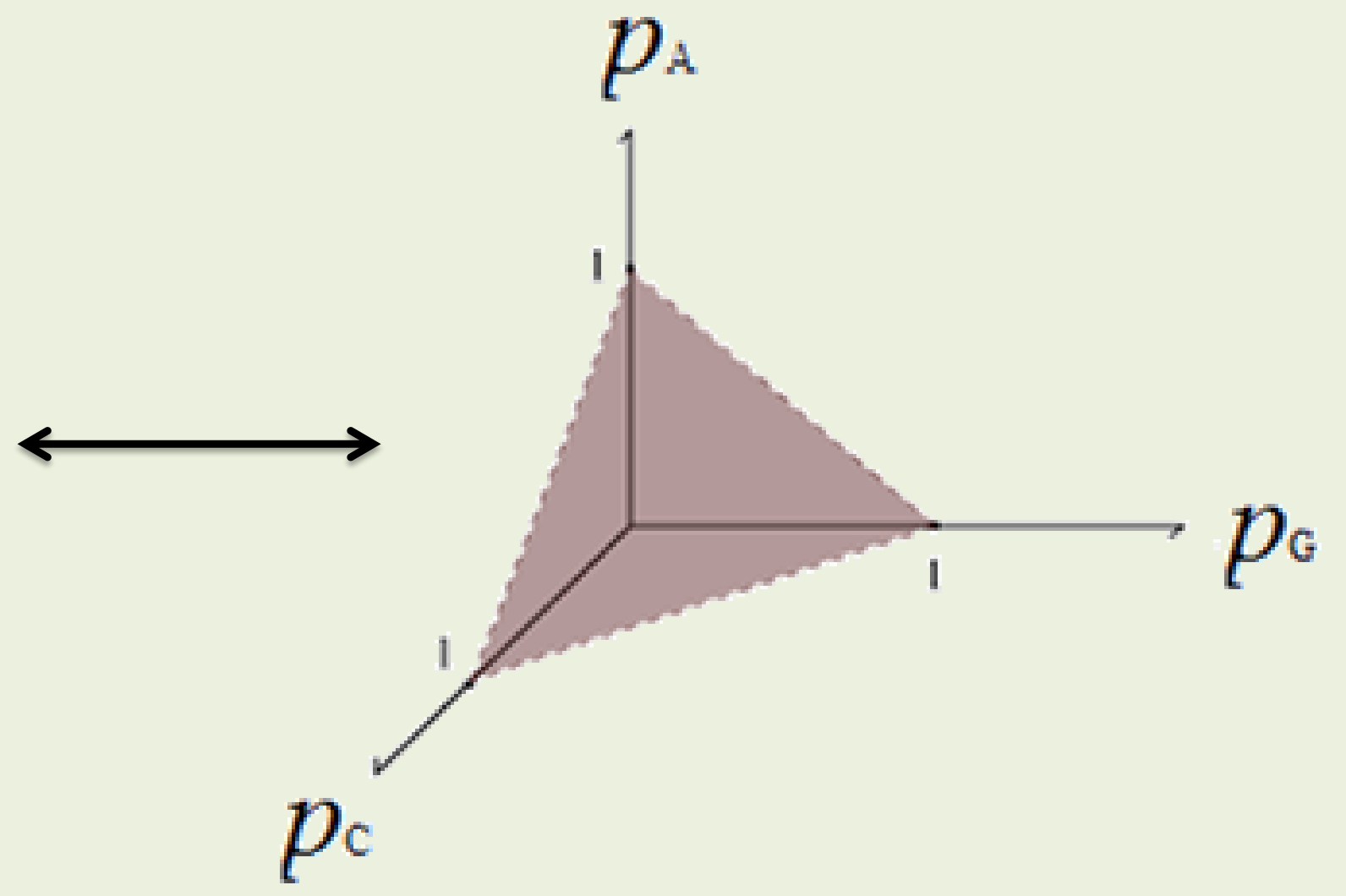
Notice that the p(seq.) above is just a polynomial function of θ_1 and θ_2. We can thus apply polynomial algebra to our study of the statistical model.

Gröbner Basis

Now, let us think about the correspondence between the algebra and the geometry.

We can take these probability distribution and embed into a geometric figure as follows:

$$\{(p_A, p_C, p_G, p_T)|\textstyle\sum = 1\} \longleftrightarrow$$



The above figure lives in $\mathbb{R}^3$. Any point on this figure (inside or in the boundary) corresponds to the probability distribution for (p_A, p_C, p_G, p_T). Varying the parameters θ_1 and θ_2 gives a subspace of the tetrahedron.

A Gröbner basis is a set of multivariate polynomials that has enticing algorithmic and algebraic properties allowing us to analyze situations like the above. Every set of polynomials can be converted into a Gröbner basis.

The theory of Gröbner is based upon the concept of ideals generated by finite sets of multiple polynomials. So, we need to know some definitions, which are listed below:

Polynomial Ring

Let R be a commutative ring. The set of formal symbols

$$R[x] = \{a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0 \mid a_i \in R, \text{ n is nonnegative integer}\}$$

is called the ring of polynomials over R in the indeterminate x.

Ideal

A subring A of a ring R is called a (two-sided) ideal of R if for every r ∈ R and every a ∈ A both ar and ra are in A.

Basis

Let F = $\{f_1, \ldots, f_s\}$ be a set of multivariate polynomials. Then the ideal generated by F_1 denoted by I = <F> is given by:

$$\{\textstyle\sum_{i=1}^{s} h_i f_i \,\vdots\, h_1, \ldots hs \in k\,[x_1, x_2, \ldots x_n]\}$$

The polynomials $f_1, \ldots, f_s$ are called a basis for the ideal they generate, and since F is finite, we say the ideal is finitely generated.

Variety

The variety of $\mathcal{F}$ is the set of all common zeros:

$$\mathcal{V}(\mathcal{F}) = \{(z_1, \ldots.z_n) \in \mathbb{R}^n : f(z_1, \ldots, z_n) = 0 \text{ for all } f \in \mathcal{F}\}$$

S-polynomial

Given two polynomials f, g ∈ $k[x_1, x_2 \ldots, x_n]$. Let J = l.c.m (leading monomial (f), leading monomial(g)). We define s-polynomial of f and g as the linear combination.

$$\text{s-poly}(f, g) = \frac{J}{\text{Leading Term (f)}} \cdot f - \frac{J}{\text{Leading Term (g)}} \cdot g$$

Gröbner basis

Let F be a finite set of polynomials. Then F is a Gröbner basis if and only if every S-polynomial has remainder zero when divided by F. When F is a Gröbner basis division always produces unique remainders which we can use as a normal form for a polynomial's equivalence class modulo the ideal generated by F. Gröbner bases also allow us to solve systems of equations and to analyze varieties in an algebraic setting.

Algorithm Buchberger

The algorithm is used to calculate Gröbner's basis. The simplest algorithm to describe, and the basis for many optimizations, is "Algorithm Buchberger" which is as follows:

Input: A polynomial set $F = (f_1, \ldots, f_n)$ that generates an ideal I.
Output: A Gröbner basis $G = (g_1, \ldots, g_t)$ that generates the same ideal I with $F \subset G$.

```
$G := F$
$M := \{\{f_i, f_j\} \mid f_i, f_j \in G$ and $f_i \neq f_j\}$
Repeat
    $\{p, q\} :=$ a pair in $M$
    $M := M - \{\{p, q\}\}$
    $S := Spoly(p,q)$
    $h := NormalForm(S,G)$
    IF $h \neq 0$ THEN
        $M := M \cup \{\{g, h\}\ \forall g \in G\}$
        $G := G \cup \{h\}$
Until $M = \emptyset$
```

Application

Since each characters are generated independently, likelihood function is given by :-

$$\mathcal{L} = p_C p_T p_C p_A p_C p_G \ldots.. p_A = p_A^{10} \cdot p_C^{14} \cdot p_G^{15} \cdot p_T^{10}$$

Now, log-likelihood is given by:-

$$\ell\,(\theta_1, \theta_2) = \log(\mathcal{L}(\theta_1, \theta_2))$$
$$= 10 \,.\, \log(p_A(\theta_1, \theta_2)) + 14.\ \log(p_C(\theta_1, \theta_2)) + 15 \,.\, \log\,(p_G(\theta_1, \theta_2)) + 10 \,.\, \log(p_T(\theta_1, \theta_2))$$

The solution to this optimization problem can be computed in closed form, by equating the two partial derivatives of the log-likelihood function to zero:

$$\frac{\partial l}{\partial\theta 1} = \frac{10}{p_A} \cdot \frac{\partial p_A}{\partial\theta 1} + \frac{14}{p_C} \cdot \frac{\partial pC}{\partial\theta 1} + \frac{15}{p_G} \cdot \frac{\partial pG}{\partial\theta 1} + \frac{10}{p_T} \cdot \frac{\partial pT}{\partial\theta 1} = 0,$$
$$\frac{\partial l}{\partial\theta 2} = \frac{10}{p_A} \cdot \frac{\partial pA}{\partial\theta 2} + \frac{14}{p_C} \cdot \frac{\partial pC}{\partial\theta 2} + \frac{15}{p_G} \cdot \frac{\partial pG}{\partial\theta 2} + \frac{10}{p_T} \cdot \frac{\partial pT}{\partial\theta 2} = 0.$$

Using Gröbner bases, we obtain

$$13003050 \,.\, \theta_1 + 2744 \,.\, \theta_2^2 - 2116125 \,.\, \theta_2 - 6290625 = 0 \ \ldots\ldots(i)$$
$$134456 \,.\, \theta_2^3 - 10852275 \,.\, \theta_2^2 - 4304728125 \,.\, \theta_2 + 935718750 = 0 \ \ldots..(ii)$$

Solving equation (i) and (ii), we get

$$(\hat{\theta}_1, \hat{\theta}_2) = (0.5191263945, 0.2172513326).$$

The log-likelihood function attains its maximum value at this point:

$$\ell(\hat{\theta}_1, \hat{\theta}_2) = -67.08253037.$$

The corresponding probability distribution

$$(\hat{p}_A, \hat{p}_C, \hat{p}_G, \hat{p}_T) = (0.202432, 0.289358, 0.302759, 0.205451)$$

is very close to the empirical distribution

$$\frac{1}{49}(10, 14, 15, 10) = (0.204082, 0.285714, 0.306122, 0.204082).$$

References

[1] B. Buchberger. A Criterion for Detecting Unnecessary Reductions in the Construction of Grobner Bases. Lecture Notes in Computer Science, vol. 72. Springer-Verlag, 1979.

[2] Pachter, Lior, and Bernd Strumfels. *Algebraic Statistics for Computational Biology*. Cambridge: Cambridge UP, 2005. Print.

[3] T. Becker and V. Weispfenning. Grobner Bases: A Computational Approach to Commutative Algebra, Springer-Verlag, 1993.