

Sequence Alignment, cont.

Statistical Significance.

There are a number of different issues, methods involved in assessing the significance of the match scores we have described. In gross terms, the methods consist of analytical and simulation approaches. Numerical simulation is most often used, but analytical examples are useful for calibration, interpretation of the simulation data. The main first problem, however, is to know what to expect of random scores as the lengths of sequences get large.

A.) Simplest case: “alignment given”. This means that we are given two sequences

$$\mathbf{a} = a_1 a_2 \dots a_n$$

$$\mathbf{b} = b_1 b_2 \dots b_n$$

The alignment is given, so we are **not** optimizing over alignment first. Our scores were given by comparison with a random positions

model, by summing n independent identically distributed (*iid*) random variables $s(a_i, b_i)$,

$$S(\mathbf{a}, \mathbf{b}) = \sum s(a_i, b_i).$$

The expectation and variance are given by

$$E(S) = nE(s(a, b)), \quad V(S) = nV(s(a, b)).$$

Note that this grows linearly with n , compared to the Erdős-Renyi result for runs of coin matches. The central limit theorem says more explicitly that

$$S \approx N(nE(s(a, b)), nV(s(a, b))).$$

B.) Case: alignment unknown.

When you have to adjust the alignment to optimize the score, you no longer have a score which is simply a sum of scores for each position as above, so the CLT doesn't apply.

One substitutes for this result a result from 1973, by Kingman:

For non-negative integers $0 \leq s \leq t$, we have $\mathbf{X}_{s,t}$, rv's, with

i) $s < t < u$ implies $\mathbf{X}_{s,u} \leq \mathbf{X}_{s,t} + \mathbf{X}_{t,u}$.

ii) the joint distribution of $\mathbf{X}_{s,t}$ is the same as that $\mathbf{X}_{s+1,t+1}$.

iii) $E(\mathbf{X}_{0,t}) = g_t$ is finite and is bounded below by $-Kt$.

Then the limit $\lim_{t \rightarrow \infty} \mathbf{X}_{0,t}/t = \lambda$ with probability 1, and in the mean.

Compare this to the SLLN. The difference here: can't say what λ is in general. This result applies to alignment scores. [See the Appendix.]

Simplest case: two letters, uniformly distributed and $s(a,b) = 0, a \neq b$, and $s(a,a) = s(b,b) = 1$. Longest common subsequence. Even here, the exact value of λ is not known.

The main point here is that optimal score grows linearly with sequence length.

C.) Local Comparisons: The Erdős-Renyi law is the first model here, and we saw that it showed logarithmic growth with sequence length. Recall that it said, considered as a LLN:

$$\lim R_n = \log_{1/p} n,$$

where p is the bias for a head and R_n is the length of the longest run of heads in the sequence of n flips of the coin. To get sharper information for estimating p -values, we use the Poisson clumping method. The first intuition is that if we are looking for a heads run of length t , it has a very low probability of occurring, but the number of places where it can *begin* is very large. This suggests that it might behave like a Poisson random variable. The heads will, for this problem, occur in clumps: if you have such a run at a given position, it is much more likely than usual to find more nearby. So, if the coin has bias p , if a run of

length t starts at the position i , say, then there is a probability p that one will start at $i + 1$, p^2 for $i + 2$ and so on. So the average clump size should be approximately $1 + p + p^2 + \dots = \frac{1}{1-p}$. The idea for a better approximation is to only count the first occurrence in every clump.

We consider a *test length* t and let \mathbf{X}_α be indicator rv's for whether or not a run of length t occurs. We want to estimate the probability that $\mathbf{W} = \sum \mathbf{X}_\alpha = 0$, that is, that the longest run is less than t . We know what this should be for a Poisson distribution, which is determined by an expectation or intensity of arrivals parameter λ . The analysis here leads to the following value for this λ :

$$\lambda_n(t) = p^t((n - t)(1 - p) + 1).$$

The test length is found by requiring λ to be bounded away from 0 and ∞ . This works out to be

$$t = \log_{1/p}((n - t)(1 - p) + 1) + c,$$

for some constant c . Putting these together gives

$$P(R_n < t) \approx e^{-\lambda_n(t)}.$$

D.) Local Comparisons, Alignment Unknown. This is the general case, and there are some results by Arratia and Waterman. Here the dependence is not so simple to elucidate, but one gets a similar formula for a constant λ so that

$$\lambda = p^t[(n + m - 2t + 1) + (n - t)(m - t)(1 - p)].$$

We will see more about these approximations in examples (the ECORRD example) in lab. It turns out that a more precise approximation is needed to see any difference in the similarity between those various tRNA's and the 16S rRNA of *E. coli*.

E.) "Phase transitions". So, we have seen that sometimes the match score is supposed

to grow linearly with n and sometimes logarithmically. It is a model result of Waterman and others that these two behaviors “suffice”.

F.) Extreme Value Statistics. Karlin-Altschul Statistics.

Review the “Erdős-Renyi heuristic”: a headrun of length m has probability p^m , but there are approximately n possible places to start this run (these are not independent, of course). So, we estimate

$$E(\# \text{ headruns of length } m) \approx np^m.$$

If we also expect that *the maximal run is unique*, then its length R_n satisfies

$$1 = E(\# \text{ headruns of length } m) \approx np^{R_n}.$$

This leads to

$$R_n = \log_{1/p} n,$$

which is the expected value of the maximal run length.

Let's consider matching according to a scoring matrix now: $s(a, b)$.

Notation: $A_1, A_2, \dots, A_n, \dots, B_1, B_2, \dots, B_n, \dots$ are iid rv's, independent, and $\mathbf{A} = A_1 A_2 \dots, \mathbf{B} = B_1 B_2 \dots$ are two random sequences built out of these letters. The distribution for the letters is going to be $p_i = \text{Prob}(A = a_i)$.

Assumptions:

(1) $\sum_{i,j} p_i p_j s(a_i, a_j) < 0$ (a random segment alignment should have negative score)

(2) $s^* = \max\{s(a, b) | a, b \text{ such that } p_a p_b \neq 0\} > 0$ (we expect to be able to find some alignments with positive score).

(3) **no gaps are allowed.**

Let H_n denote the maximal score for an un-gapped local alignment of subsequences of

$A_1A_2 \dots A_n$ and $B_1B_2 \dots B_n$. The KA result says that this will behave like ER. What bias?

Let p be the (unique) positive solution of

$$1 - E(p^{-s(a,b)}) = 0 = 1 - \sum p_a p_b p^{-s(a,b)}.$$

KA1:

$$P\left(\lim_{n \rightarrow \infty} \frac{H_n}{\log_{1/p}(n)} = 2\right) = 1.$$

This should tell us how to normalize the max-score if we can get an EVS for it.

Interesting: the distribution of letters in the optimal aligning segments can be *different* from p_a . Let it be given by $q_{a,b}$. Then one has

KA2

$$p_{a,b} = p_a p_b p^{-s(a,b)},$$

and the KA condition selecting p is exactly the condition guaranteeing that $q_{a,b}$ is a probability distribution.

Remark on the “bias”: we can rewrite our condition for p as a condition on λ in the equation

$$\sum p_a p_b e^{\lambda s(a,b)} = 1,$$

where $\lambda > 0$ (so that $p = e^{-\lambda}$ satisfies $0 < p < 1$). Now if we look at our optimal alignment and score algorithms, the optimal alignment is unchanged if we replace $s(a,b) \rightarrow \lambda s(a,b)$, for any $\lambda > 0$. (All scores will simply *scale* by λ .) Thus, the theorem says there is a natural scale for a scoring matrix.

Also, when we made up scoring matrices from probabilities as we did, we found

$$s(a,b) = \log \frac{p_{a,b}}{p_a p_b}.$$

This scoring matrix already *is* in “natural scale” (i.e., $\lambda = 1$ for this scoring matrix already):

$$\sum p_a p_b e^{s(a,b)} = \sum p_a p_b \frac{p_{a,b}}{p_a p_b} = \sum p_{a,b} = 1.$$

Erdős-Renyi heuristic for KA: consider $\mathcal{A} \times \mathcal{A}$, the pairs of letters, and let Q be a distribution on it, and $R = p_a p_b$ the product distribution. There are about n^2 places to begin a matching segment of length k . Such a sequence would on average

(i) score $\sum s(a, b)q_{a,b}$ per letter

(ii) occur with probability approximately $e^{-kH(Q||R)}$, where $H(Q||R)$ is the relative entropy.

This last uses a “large deviation” result as motivation. See below in the appendix.

ER suggests solving

$$n^2 e^{-kH(Q||R)} = 1,$$

giving

$$k \approx \frac{\log n^2}{H(Q||R)}.$$

Knowing the average score per letter and the average length of the best aligned segment for

the distribution Q , the best alignment score should be

$$\frac{\sum q_{a,b} s(a,b)}{H(Q||R)} \cdot \log n^2.$$

Now we can vary Q , and try to maximize this score over the distributions, that is, maximize

$$r(Q) = \frac{\sum q_{a,b} s(a,b)}{H(Q||R)}$$

over possible distributions Q on $\mathcal{A} \times \mathcal{A}$. Since we are constrained to have $\sum q_{a,b} = 1$, we use Lagrange multipliers and set

$$\frac{\partial r(Q)}{\partial q_{a,b}} = \lambda \frac{\partial \sum q_{a,b} - 1}{\partial q_{a,b}} = \lambda.$$

That is,

$$\frac{s(a,b)}{H(Q||R)} - \frac{1 + \log \frac{q_{a,b}}{p_a p_b}}{H(Q||R)^2} = \lambda.$$

Solving as we did with the Boltzmann distribution some days ago, we get

$$\log \frac{q_{a,b}}{p_a p_b} = c s(a,b) + c_1,$$

or

$$q_{a,b} = \frac{e^{cs(a,b)}}{c^*} p_a p_b = \frac{e^{cs(a,b)} p_a p_b}{\sum_{a,b} p_a p_b e^{cs(a,b)}}.$$

where c, c_1 are constants. We can show that when $r(Q)$ is maximal, then $D(c) = \sum p_a p_b e^{cs(a,b)}$ is just 1. Do this by setting

$$\frac{dr(Q(c))}{dc} = 0.$$

KA show that the maximum score has an approximately EVS distribution. Specifically, they find that if we consider *ungapped, local optimal alignment* between random sequences **A**, **B** of lengths m, n respectively, then the number of alignments with score $\geq S$ is \approx Poisson with mean $E(S) = K m n e^{-\lambda S}$, where $\lambda > 0$ gives the “natural scale”

$$\sum p_a p_b e^{\lambda s(a,b)} = 1,$$

and K is given by an explicit (infinite) sum. Finally, this gives us

$$P(x > S) \approx 1 - e^{-E(S)}.$$

The constants λ, K have been compiled by Altschul and others for a variety of standard scoring matrices.

Many simulation tests have been run on the mathematically less tractable case of alignments with gaps allowed, and the results are similar to those for ungapped alignment. There is no analytical calculation in this case, however.

G. Information Interpretation of Scoring Matrices.

An obvious question: which scoring matrix is appropriate to use? Altschul (and others) addressed this. Altschul observed that, given the implicit “target frequency” $q_{a,b}$ associated to a similarity scoring matrix $s(a,b)$, one could scale the matrix beforehand so that the λ in

$$1 = \sum p_a p_b e^{\lambda s(a,b)}$$

is just $\log 2$. Then the expected score at each (independent) position in the optimal scoring

segments, which are distributed according to the *target distribution* Q , will be given by

$$\sum q_{a,b} \log_2 \frac{q_{a,b}}{p_a p_b},$$

the relative entropy measured in *bits*. So, the score of a sequence alignment can be thought of in bits, and given independent positions, these bit scores are expected to add with the number of positions (HW exercise).

Suppose we have p segment pairs with scores $\geq S$. Then the EVS gives $p \approx K m n s^{-\lambda S}$ where λ is now \log_2 . Taking log's base 2, we get

$$S \approx \log_2 \frac{K}{p} + \log_2 m n$$

. Altschul finds that for the typical scoring matrices, K is relatively stable, usually ≈ 0.1 . Taking $p = 0.05$ to see what the best score might look like, one would get that S was, for large $m n$ approximately $\log_2 m n$. This says the score is basically containing the information of where to start the two segments which are matched (aligned).

On the other hand, one expects $H(Q||P)$ bits per position of relative information against the random background. As evolutionary time progresses, one expects the behavior at a given position to be randomized, and that is indeed what happens. That is, as the time index of a PAM matrix gets large, the $H(Q||P)$ gets smaller, each position in an alignment offers less information against the random background. Thus, to get a significant score, an alignment with a higher PAM time parameter must be longer. The graphs are presented here. The main content of the graphs is what the entropy evaluations of the various PAM's are. The claim then was that only PAM 40, 120, 250 were essential, though others are still used. Similar remarks apply to BLOSUM matrices.

H. An RNA Example.

This is an example already seen in outline by the students in the lab. It concerns the relations between transfer RNA and ribosomal RNA. The data is drawn from *E. coli*.

Recall that the ribosome is a large complex consisting of a mixture of much RNA and several proteins holding the RNA components together. It is the locus of protein synthesis in cells, and is remarkably conserved in the *RNA secondary structure* across species. (More on this later in the term.) There are two main RNA components of the ribosome, and they are classified according to a (sedimentation) measure of their weight. The data we are going to look at includes the smaller 16S component of *E. coli* rRNA.

Proteins are assembled linearly along the chain, adding one amino acid at a time. The successive amino acids are brought to the correct position for incorporation into the growing chain by an RNA molecule, transfer RNA (tRNA). There are over twenty of these molecules, each one specific for a given amino acid.

A great question in biology is the origin of life, or more modestly, at least what was the step

in the evolution to the current form of life that preceded us? The gist of the problem is: one needs to carry information for reproduction (genetics; nucleic acids), while it is also clear that one needs energy to perform the functions necessary for carrying out this reproduction (metabolism, and more generally, enzymatic activity; proteins). Is it possible to decide whether proteins came first or nucleic acids, or are there fossil remains of some other chemical basis for the mechanics of life?

A decisive step was taken around 1980 when Cech and coworkers showed that RNA could work as a catalyst as well as an information molecule. Since then there is a growing body of observation supporting the so-called RNA world hypothesis. The most recent support came from x-ray crystallography work of the groups of H. Noller at UCSC and T. Seitz at Yale on the details of the ribosome. It was found, most importantly from this point of view, that the active sites of synthesis in the ribosome consisted entirely of rRNA, that is, the

many proteins in the complex were only there for structural reasons, they did not perform the fundamental chemistry.

One of the first attempts to follow on the discovery of Cech et al. was to see whether there was an evolutionary arrow pointing from tRNA to the ribosome, to argue for a direction of evolution, presumably from simpler to more complex molecules. There was some interest in the first results in this direction, which were among the earliest uses of sequence comparison to search for evolutionary relationships. Unfortunately, there was some “irrational exuberance” in the estimation of the significance of the matches found. The data presented in the first figure (from a review article of Waterman from 1995) is looking at the statistics of the scoring function, the “longest exact match”. This falls into the cadre of the Poisson approximation outlined earlier, and the data includes the name of the tRNA species (there are some redundancies), the length of the molecules, the

longest exact match, p -value and " b_1 ", which is a measure of the overlap (non-independence) of the long match runs. (Smaller b_1 means closer to being independent.)

The basic conclusion is that the matches are none of them significant in any meaningful sense!

This does not mean that the biological hypothesis is wrong. As noted, the remarkable thing about the whole ribosomal mechanism is its great age and enduring stability. One estimate is that the sequences we are looking at originated about 3 billion years ago. Thus a much better idea would be to look at local alignment with a score that allowed for mutations, etc. The second data example shows exactly this. The statistics are those for the SW alignment with $+1$ for a match position, the penalty $\mu = 0.9$ for a mismatch, and the gap penalty (linear) is $\delta = 2.1$. Unfortunately, in this display, the significance is calculated in standard deviations, so that $\# \sigma$ means the

number of standard deviations above the mean for the scoring statistic. These are evaluated by simulation calculations. The formula for the max score is approximately

$$H \approx 5.04 \log n - 30.95,$$

while the sample variance $\hat{\sigma} \approx 1.49$. Notice that cysteine tRNA is now seen to have a radically different significance level. There is no biological explanation of this yet.

Appendix: Some Calculations and Proofs.

1.) Comparison of Kingman's theorem with the SLLN. To compare these, we are given a sequence of iid's $\mathbf{X}_1, \dots, \mathbf{X}_n$, and we set $X_{s,t} = \mathbf{X}_{s+1} + \dots + \mathbf{X}_t$. Then

$$\mathbf{X}_{s,u} = \mathbf{X}_{s,t} + \mathbf{X}_{t,u}, 0 < s \leq t \leq u.$$

The other hypotheses of Kingman's theorem are easily verified, for example,

$$E(\mathbf{X}_{0,t}) = tE(\mathbf{X}_1) \geq -t|E(\mathbf{X}_1)|.$$

Kingman's theorem only tells you that

$$\lim \frac{\mathbf{X}_{0,t}}{t} = \lambda,$$

with probability 1, for *some* λ , a constant. The SLLN says $\lambda = E(\mathbf{X}_1)$.

A proof of Kingman's theorem can be found in the book of Durrett, "Probability: Theory and Examples", which contains a number of other results we are looking at.

2.) Derivation of the Extreme Value Distribution.

Suppose we are given a random variable \mathbf{X} and a sequence of iid rv's $\mathbf{X}_i \sim \mathbf{X}$. Consider the rv $\mathbf{M}_N = \max\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$, and we want to calculate the distribution for \mathbf{M}_N as N gets large. The standard argument uses the distribution function

$$G(x) = Prob(\mathbf{X} \leq x) = \int_{-\infty}^x g(s) ds,$$

and seeks the p.d.f. for \mathbf{M}_N as

$$\frac{d}{dx} \text{Prob}(\mathbf{M}_N \leq x) = \frac{d}{dx} G^N(x) = N g(x) G^{N-1}(x).$$

If we assume that $\mathbf{X} \sim \text{Exp}[\lambda]$, then $g(x) = \lambda e^{-\lambda x}$, $G(x) = 1 - e^{-\lambda x}$. We have

$$N g(x) G^{N-1}(x) = N \lambda e^{-\lambda x} (1 - e^{-\lambda x})^{N-1}.$$

The key is to center the distribution, that is, to subtract the right constant c_N from \mathbf{M}_N so that the limit will exist. This amounts to having a guess as to the mean for the max distribution. In this case we choose c_N so that $e^{-\lambda c_N} = \frac{1}{N}$, or

$$c_N = \lambda^{-1} \log N.$$

One makes this choice precisely to eliminate the factor N in front of the last distribution. Setting $z = x - c_N$, and rewriting the above distribution in terms of z , we get:

$$N \lambda e^{-\lambda x} (1 - e^{-\lambda x})^{N-1} = N \lambda e^{-\lambda z - \lambda c_N}$$

$$\begin{aligned} & \times (1 - e^{-\lambda z - \lambda c_N})^{N-1} \\ & = \lambda e^{-\lambda z} (1 - e^{-\lambda z / N})^{N-1}. \end{aligned}$$

Now, if z were fixed, then

$$\begin{aligned} \lim_{N \rightarrow \infty} (1 - e^{-\lambda z / N})^{N-1} &= \lim_{N \rightarrow \infty} (1 - e^{-\lambda z / N})^N \\ &= e^{-e^{-\lambda z}}, \end{aligned}$$

a double exponential, which goes to 1 as $z \rightarrow \infty$ very rapidly. We are using the basic fact: $\lim (1 + \frac{a}{N})^N = e^a$, for any number a . However, we also have to worry about all z at once. This proof is not complete: we have to show that the probability that z is very large is small enough to make this argument work for all z at once.

Returning to the main argument, we have the p.d.f. of $\mathbf{M}_N - c_N$ is given by

$$\lambda e^{-\lambda z} e^{-e^{-\lambda z}}.$$

This gives us the approximation

$$Prob(\mathbf{M}_N - c_N \geq a) \approx 1 - Prob(\mathbf{M}_N \leq a)$$

$$\approx 1 - \int_{-\infty}^a \lambda e^{-\lambda z} e^{-e^{-\lambda z}} dz = 1 - e^{-e^{-\lambda a}}.$$

This distribution is also known as a *Gumbel* distribution. Surprisingly, a wide family of \mathbf{X} 's will in fact have the same sort of EVS, and in general, there are only a very small number of EVS which arise (3; Gumbel is also sometimes called a “type I EVS”).

Here is the general result on EVS of Type I:

$\mathbf{X}_1, \dots, \mathbf{X}_n, \dots$ are iid rv's, with common distribution function $G(x)$.

$$\mathbf{Y}_N = \max\{\mathbf{X}_1, \dots, \mathbf{X}_N\}.$$

Assume there are sequences a_N, b_N so that

$$\lim_{n \rightarrow \infty} N\{1 - G(a_N + b_N y)\} = u(y)$$

exists for all y . Then

$$\lim_{N \rightarrow \infty} P(\mathbf{Y}_N < a_N + b_N y) = e^{-u(y)}.$$

To compare to the case just treated, a_N would be $c_n = \frac{\log n}{\lambda}$, and b_N would be 1, giving $u(y) = e^{-\lambda y}$.

3. Large Deviations.

We wanted an approximation to the probability of finding, in a pair of random sequences \mathbf{A}, \mathbf{B} , a segment of length k where the distribution is given by $Q = \{q_{a,b}\}$. The point is that the distribution Q has a higher average match score per position, and so such sequences will raise the score above what one would randomly expect. The claim is that for the optimal scoring segments, the probability of finding one of length k will be about $e^{-kH(Q||R)}$ as above.

We will only treat here explicitly the following simplified coin flip case: let \mathbf{Y}_n be the rv “the number of heads in n independent flips of a coin with bias p ”. Think of this as a match score which counts the total number of exact matches with the all heads sequence. Consider

a number α with $p < \alpha < 1$. Let Q be the distribution $(\alpha, 1 - \alpha)$ for coin flips, and $P = (p, 1 - p)$ the original one. Note Q has higher expectation. The elementary result is:

$$\text{Prob}(\mathbf{Y}_n \geq \alpha n) \leq e^{-kH(Q||P)},$$

where $H(Q||P)$ is the relative entropy.

This is easy to see:

$$\text{Prob}(\mathbf{Y}_n \geq \alpha n) = \text{Prob}(e^{c\mathbf{Y}_n} \geq e^{c\alpha n}),$$

where $c > 0$ is any constant.

$$\text{Prob}(e^{c\mathbf{Y}_n} \geq e^{c\alpha n}) \leq E(e^{c\mathbf{Y}_n})/e^{c\alpha n},$$

by Markov's inequality.

$$E(e^{c\mathbf{Y}_n})/e^{c\alpha n} = \{e^{-c\alpha}(1 - p + pe^c)\}^n.$$

Use calculus to minimize the right hand side ($\frac{d}{dc} = 0$), and get

$$c = \log \frac{\alpha}{1 - \alpha} \frac{1 - p}{p} = \log r,$$

where r is the “odds-ratio” between a p -coin and an α -coin. From this you see that the best you can do is

$$e^{-\alpha \log r + \log(1-p)} + e^{(1-\alpha) \log r + \log p}.$$

It is not so easy to see that this is $e^{-H(Q||P)}$. You can think of α as a parameter, and notice that the two agree when $p = \alpha$. For the rest compare derivatives using

$$\frac{dH(Q||P)}{d\alpha} = \log r.$$

In using the large deviation estimate in the ER heuristic, we think of k as being the length of a maximal scoring subsequence, so it makes sense to substitute

$$\text{Prob}(\text{length of segment is } k) = e^{-kH(Q||P)}.$$

4. Arratia-Waterman Phase Transition: An Example.

We consider the local alignment problem of two sequences from an alphabet \mathcal{A} with

$$s(a, b) = \begin{cases} +1, & a = b \\ -\mu, & a \neq b \end{cases}$$

and linear gap penalty given by $-\delta$ for each gap position. Then we have a family of problems given by $(\mu, \delta) \in [0, \infty]^2$. If

$$S_N = S(A_1 A_2 \dots A_N, B_1 \dots B_N),$$

we know

$$\lim E(S_N)/N = \rho$$

exists. There is curve in the (μ, δ) quadrant where $\rho = 0$ separating where $\rho < 0$ from $\rho > 0$. Basically, the result is that for $\rho < 0$, the best score H_N grows like $\log N$, while for $\rho > 0$, H_N grows linearly. See if you recognize where our special cases are on the boundary of the quadrant.

References.

The best single place reference for the statistics of scoring and alignments is Chapter 11 of

Michael Waterman's, "Introduction to Computational Biology", referred to in the original course announcement. Unfortunately, there are several issues intermingled in this chapter, and parts are difficult. The original papers develop through the '80's, culminating to some degree in the paper by Karlin and Altschul in *PNAS* **87**, 1990, pp. 2264-2268, though the full arguments required later papers of Karlin and other coworkers.