1 Global versus local alignments

Thus far we have been finding alignments over the entire sequence, yet typically biologists think alignment over short distances is what really matters for functional conservation in many instances. The alignments we have sought are known as *global alignments*, in contrast to aligning short segments of 2 given sequences, which is known as *local alignment*. Local alignment is more relevant because many proteins have functions that are mediated by their ability to bind to another molecule (the *ligand* of the protein), hence function will be preserved if this shorter segment is retained through evolution even if there is substantial divergence in other regions of the proteins. Since proteins are folded in their natural state, these preserved regions needn't actually be continuous segments of the protein. In fact, many researchers working on lymphocyte recognition of antigen explicitly account for these discontinuities in binding domains (so called "non-linear" epitopes, where an *epitope* is the ligand of a lymphocyte).

We can easily modify our global alignment algorithm to become a local alignment algorithm by introducing stops in our dynamic programming algorithm. That is, when we compute the maximal scores at each location we add a fourth option: have no alignment. This means instead of just comparing the 3 possible scores, we add another option in which the total score is simply zero. If $F(i,j)$ represents the value in the array of scores, g is the gap penalty, r_{ij} is the j^{th} element of the i^{th} sequence, then we now have

$$F(i,j) = \max(0, F(i-i,j)+g, F(i,j-1)+g, F(i-1,j-1)+s(r_{1j}, r_{2j})),$$

where s is the score for 2 elements from the sequences. With this modification, one then proceeds to fill in the matrix $F(i,j)$ as before. To find the optimal local alignment using this matrix of scores, one first determines the minimal value in the matrix. This is where the optimal local alignment *ends*. One then traces back from this location to the first location which is zero in the matrix F. This point is the start of the optimal local alignment. This algorithm is known as the *Smith-Waterman algorithm*.

2 Statistical significance of sequence similarity

2.1 Evolutionary Models and testing the hypothesis of no relation

The basis for sequence alignment is that there is a common ancestor for the sequences under investigation. We derive each of the 2 sequences under comparison by a sequence of mutations, insertions and deletions (insertions and deletions are referred to as *indels*). The process of obtaining a sequence from another by these basic operations works forward and backward in time, i.e. a deletion forward in time is an insertion if one reverses time. Hence, if a_0 is the common progenitor of the 2 sequences, a and b, our evolutionary model postulates that $a = M_s(M_{s-1}(\cdots(a_0)\cdots)$ and $b = M_t(M_{t-1}(\cdots(a_0)\cdots)$, where M_s represents a mutation or indel. But since mutations and indels are time reversible, we can write

$$a = f_J(f_{J-1}(\cdots(b)\cdots),$$

for some sequence of mutations and indels, f_j. In this framework, the hypothesis that the 2 sequences are not related through evolution is $H_0 : J \geq J_0$, where J_0 is large enough so that the protein has lost its function.

The practical approach to statistical significance adopted is to use data on proteins with known evolutionary similarity and use these to determine the likelihood of a given mutation. As discussed in the section on finding the most likely alignment, we can use expert knowledge to obtain estimates of given mutations, and then use dynamic programming to find the most likely alignment and an alignment score. In fact, the alignment score is the log of the likelihood ratio for testing if the 2 sequences are independent against

the alternative that they are related and the probabilities for mutations are as specified in the substitution matrix (and gap penalty) being used. The natural question then is if this alignment score is extreme enough to conclude the alignment is not just chance agreement. Unlike the usual applications of the theory of maximized likelihood ratio statistics, the number of parameters increases with the number of independent observations on the process (i.e. the length of the 2 sequences) since the alignment itself is a parameter that is being estimated from the data. This makes the usual theory of such tests not relevant, and indeed the distribution of the alignment score under the null is certainly not like any χ^2 random variable.

Note that the optimal alignment is the highest of many possible scores, that is, if there are N distinct paths through the lattice introduced for finding the optimal alignment, and if s_i is the score associated with the i^{th} path, then our alignment score is $S = \max_i s_i$. Since many of the paths necessarily intersect, the scores themselves s_i can not be independent, although this dependence is weak if the length of the sequences is long. In addition, we can model these scores as drawn from the common distribution defined by the sets of sums of gap penalties and alignment scores. Hence to determine the distribution of the alignment score under the null hypothesis, we need to consider the distribution of the maximum of a set of i.i.d.. random variables.

2.2 Distributions of maxima of sets of i.i.d.. random variables

In a paper of Gnedenko (1943) it was shown that the asymptotic distribution of the maximum of a set of independent and identically distributed variables has one of only 3 possible distributions. The asymptotic distribution depends on the tail behavior of the cdf. We will consider only the 2 cases where the random variable is not bounded. Let $Y_{(k,n)}$ represent the k^{th} order statistic based on a sample of size n. Recall that the cdf of the maximum of set of i.i.d. variables with cdf F is $F(x)^n$. If F represents the cdf and $1 - F(y) \sim ay^{-c}$ for some a, c (both positive) then for any sequence b_n

$$\begin{aligned} P(Y_{(n,n)}/b_n \leq z) &= \Big[1 - \Big(1 - F(b_n z)\Big)\Big]^n \\ &\sim \Big[1 - a(b_n z)^{-c}\Big]^n. \end{aligned}$$

So letting $b_n = (an)^{1/c}$, we find

$$P(Y_{(n,n)}/b_n \leq z) \sim \mathrm{e}^{-z^{-c}}.$$

If $1 - F$ goes to zero exponentially, i.e. $1 - F(y) \sim \mathrm{e}^{-y}$ then we obtain a different distribution for the maximum. In this case

$$P\Big(\frac{Y_{(n,n)} - a_n}{b_n} \leq z\Big) = \Big[1 - \exp\Big\{\log\Big(1 - F(a_n + b_n z)\Big)\Big\}\Big].$$

In this case we expect a_n to be near $F^{-1}(1)$, and we will set $a_n = F^{-1}(1 - 1/n)$. Let

$$g(z) = \log\Big(1 - F(a_n + b_n z)\Big).$$

Now Taylor expand g near $z = 0$ (using $a_n = F^{-1}(1 - 1/n)$) to obtain

$$g(z) = \log(n^{-1}) - n b_n f(a_n) z,$$

where f is the density of the random variables, so that

$$\begin{aligned} P\left(\frac{Y_{(n,n)} - a_n}{b_n} \leq z\right) &= \left[1 - \frac{1}{n}\exp\Big\{-nb_n f(a_n)z\Big\}\right]^n \\ &\sim \left[1 - \frac{1}{n}\exp\{-z\}\right]^n, \end{aligned}$$

if we set $b_n = 1/(nf(a_n))$. But then taking the limit in n we find

$$P(\frac{Y_{(n,n)} - a_n}{b_n} \leq z) \sim \exp\{-\mathrm{e}^{-z}\}.$$

This distribution is known as the *extreme value* distribution. Note that the above argument implies $\mathrm{E}Y_{(n,n)} = a_n + c$, but $a_n \sim \log n$, so $\mathrm{E}Y_{(n,n)} \sim \log n$.

The usual approach supposes the tail area of the distribution of alignment scores decays at an exponential rate, hence the extreme value distribution is used to determine tail areas. This is appropriate since the scores are sums and provided the correlation within a sequence isn't too strong a central limit theorem implies the convergence of these sums to a normal distribution. When we are finding the local alignment, the size of the largest local alignment will depend on the product of the length of the sequences nm, hence the sequence alignment score, S, is such that $S \approx \log nm + C$ or $S \approx \log Knm$. For this reason, we typically parameterize the extreme value distribution as $\exp\{-Knm\mathrm{e}^{-\lambda S}\}$ for 2 parameters K and λ (here S is the alignment score). Since not all of the sequences in the database are actually the same size, the scores are not really i.i.d. but this is usually ignored. In addition, to calculate tail areas, we need estimates of K and λ. These have been obtained empirically for a number of scoring schemes. We refer to the resulting p-values as e-values.

3 Rapid methods of sequence alignment

Often a researcher has a sequence and wants to find the sequences in a large database that are similar to the new sequence. When the sequences are 1000s of base pairs long, dynamic programming can take a minute. Thus if we are searching through a data base with millions of sequences, dynamic programming can become impractical. For this reason, a number of potentially suboptimal but rapid methods for sequence alignment are in use.

Recall from the dynamic programming formulation that if 2 sequences have extensive much agreement, the optimal path in the dynamic programming table will be largely diagonal. Hence a quick way to gauge the similarity of 2 sequences is to count the length of the longest diagonal in the dot matrix form. While it would seem summing the diagonals of a matrix would take on order nm, there are clever algorithms for computing these sums more rapidly. This is the basic idea behind the 2 most popular rapid methods for sequence alignment, BLAST and FASTA.

One very popular heuristic algorithm is known as BLAST. This algorithm proceeds by looking for short matches, known as seeds, and then tries to extend these seeds out in both directions. There are constant refinements to this *ad hoc* method. For example, recently it was proposed to speed up BLAST by looking for 2 nearby seeds and requiring them to shorter.

3.1 Bioinformatics on the web

There are many resources available on the internet. The first place to look is at the web site of the National Center for Biotechnology Information `http://www.ncbi.nlm.nih.gov/`. This page is constantly changing,

but it is a searchable site. Information on sequences is stored by *accession numbers*. If one knows the accession number, one can search for information on the sequence. For example, if you enter the number af102503, then you get one search result (under EST in the table). If you follow the link then you get a page full of information about that sequence, including the nucleotide sequence. You will see that someone from the University of Minnesota submitted this entry, it is an EST and some other information. If one goes to the bottom of the screen then there is a link to an article in *Virology* from which one could find out more abut this sequence.

One can also access the program BLAST from this website. To do this, either search for BLAST or just look for a link to BLAST (such links are frequently available in the site). There is a BLAST homepage (`http://www.ncbi.nlm.nih.gov:80/BLAST/`) that has many options. Lets look at nucleotide BLAST-so follow the link there (i.e. standard nucleotide BLAST). This brings up a screen with a blank window into which you can paste a query sequence. There are a variety of options available on this page. After submitting a query sequence, you will instructed to wait a short while, after which you get a list of the most closely matching sequences along with their alignment scores and other statistics (including e-values).

References

Gnedenko, B. (1943), "Sur la distribution limite du terme maximum d'une serie aleatoire", *Ann. Math.*, 44, 423–453.

Exercises

1. Find the distribution of the maxima of a set of independent uniform on (0,1) random variables. (*Hint* Consider $P(\frac{1-Y_{(n,n)}}{b_n} \leq x)$.)
2. For the sequence with accession number given above, what are the 2 closest matches and what are the associated e-values (be sure to search "others" in the database selection area)?