MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 10, No. 2, October, pp. 264–266, 1998
ARTICLE NO. FY980522

NOTE
The Number of Multiple Alignments

Joseph B. Slowinski

Department of Herpetology, California Academy of Sciences, Golden Gate Park, San Francisco, California 94118-4599
E-mail: jslowins@cas.calacademy.org

Received October 8, 1997; revised February 15, 1998

The following shows one possible alignment for five DNA sequences, each five nucleotides long:

AGAC---C

AGAA---C

GAAT---C

---TGACC

---TCAGC.

How many possible alignments do you think there are for five DNA sequences of five nucleotides each? Hundreds, thousands, millions? In fact, there are 1.05×10^{18} different alignments! And this is only for several, very small sequences, much fewer and shorter than those used by molecular phylogeneticists. Waterman (1994) has commented on how fast the number of alignments grows with the number and length of molecular sequences, a problem that is perhaps even more dramatic than the growth in the number of phylogenetic trees as the number of taxa increases (Felsenstein, 1978). Griggs *et al.* (1990) derived an asymptotic equation for the number of alignments. In this note, I present exact equations for the number of alignments and then argue that, because of the huge number of possible alignments for a set of sequences, strategies other than multiple sequence alignment need to be developed and implemented for establishing positional homology.

Stanton and Cowan (1970) were interested in the following recursion on positive integers: $f(n_1, n_2) = f(n_1 - 1, n_2) + f(n_1 - 1, n_2 - 1) + f(n_1, n_2 - 1)$, with the initial condition $f(n_1, n_2) = 1$ for n_1 and/or $n_2 = 0$. They showed that the recursion is solved by the equation:

$$f(n_1, n_2) = \sum_{i=0}^{n_1} \binom{n_1}{i} \binom{n_2 + i}{n_1}.$$

Although Stanton and Cowan's work was not done in the context of sequence alignment, Laquer (1981) pointed out that their recursion counts the number of alignments for two sequences of n_1 and n_2 nucleotides (or amino acids). This can be seen by considering that any possible alignment for two sequences can have only one of three endings (Waterman, 1994),

$$\ldots N \quad \ldots N \quad \ldots —$$

$$\ldots — \quad \ldots N \quad \ldots N,$$

corresponding to $f(n_1 - 1, n_2)$, $f(n_1 - 1, n_2 - 1)$, and $f(n_1, n_2 - 1)$, respectively. Stanton and Cowan's recursion cannot be generalized to multiple sequences (>2) without a huge profusion of terms ($2^m - 1$ terms for m sequences). Thus, an alternate strategy is needed. Below, I present exact equations for counting the number of alignments based on the principle of inclusion–exclusion (Roman, 1986).

Consider representing an alignment as a matrix wherein each nucleotide is represented as a 1 and each gap as a 0. Counting the number of sequence alignments is the same as counting the number of distinct matrices, subject to the following conditions: the m row sums (n_j) are held constant, any matrices with a column sum of 0 (corresponding to a column of gaps) are excluded, and the ordering of the nucleotides in a sequence is maintained (Griggs *et al.*, 1990). The number of columns can range from $N = \max(n_j)$ to Σn_j. Consider row j in one of the matrices: every one of the $\binom{N}{n_j}$ permutations of the 1s and 0s for row j will be represented at least once among the total set of possible alignments. Thus, an obvious strategy for counting the total number of alignments of a fixed size N is to take the product of the number of permutations for a single row over all rows. However, this will overcount the true number of alignments by also counting alignments with column sums of 0. Fortunately, the principle of inclusion–exclusion (Roman, 1986) can be used to sub-

1055-7903/98 $25.00

TABLE 1

A Compilation of f(n, m) for $1 \leq n \leq 5$, 10, and $2 \leq m \leq 5$

m	2	3	4	5
$n = 1$	3	13	75	541
2	13	409	23917	2244361
3	63	16081	10681263	14638756721
4	321	699121	5552351121	117629959485121
5	1683	32193253	3147728203035	1.05×10^{18}
10	8097453	9850349744182729	3.32×10^{26}	1.35×10^{38}

stract off the spurious alignments. Once this is done, it is still necessary to sum over all possible values of N ($\max(n_j) \leq N \leq \Sigma n_j$). This leads to the following equation:

$$f(n_1, n_2, \ldots, n_m) = \sum_{N=\max n_j}^{\Sigma n_j} \sum_{i=0}^{N} (-1)^i \binom{N}{i} \prod_{j=1}^{m} \binom{N-i}{N-n_j-i}.$$

If every sequence is the same length, then the equation becomes

$$f(n, m) = \sum_{N=n}^{mn} \sum_{i=0}^{N} (-1)^i \binom{N}{i} \binom{N-i}{N-n-i}^m.$$

Table 1 is a compilation of $f(n, m)$ for $1 \leq n \leq 5$, 10, and $2 \leq m \leq 5$. The reader might wonder why values for $f(n, m)$ at greater values of n and m were not tabulated. Using Maple V on a Power Macintosh 8600/200, I was unable to calculate $f(n, m)$ for values of n and m greater than those in Table 1; the numbers are simply too large.

One might argue that many if not most of the alignments in the set of all possible alignments for any pair of values n and m are highly unrealistic. For example, most molecular phylogeneticists would probably agree that the following alignment

```
AGAC---C
AGAA---C
GAAT---C
---TGACC
---TCAGC
```

is more realistic than one with more gaps, such as the following:

```
A---GAC---C
A---GAA---C
--GAA--T---C
---TG--AC-C
------TCAGC.
```

But restricting attention to alignments with smaller gaps does not eliminate the problem of the huge number of alignments. For example, if we count the number of alignments of five sequences each with five nucleotides subject to the constraint that each sequence contain no more than three gaps (done by counting from $N = 5$ to $N = 8$), $f(n, m)$ is still over 500 million!

This note has shown that the number of possible alignments for multiple molecular sequences is staggering. The implication for molecular phylogenetics is simple: computer programs that attempt to find the best alignment under some criterion of optimality are virtually guaranteed to fail, unless the data have been relatively free from insertions and deletions.[1] This, coupled with the fact that the optimal alignment is not likely to be the correct one, ensures that some positions in a molecular data set will be misaligned. One wonders how much error is introduced into molecular phylogenetic analyses from incorrect positional homology. It is clear that better strategies than the common two-step procedure of first finding an optimal multiple sequence alignment followed by phylo-

[1] Protein-coding genes generally experience fewer insertions and deletions than non-protein-coding genes, such as rRNA genes; the comments in this section, therefore, are more applicable to non-protein-coding genes.

genetic analysis of the aligned sequences need to be found. Wheeler (1996) describes a method called optimization alignment that seeks to find the most parsimonious tree without first having to perform sequence alignment. His method assigns ancestral sequences to interior nodes without inserting gaps between nucleotides, which makes sense because gaps do not constitute observations (Wheeler, 1996). The algorithm does involve alignment of each pair of sequences assigned to the two immediate descendant nodes of each interior node, but this is a vast improvement over having to consider all possible multiple sequence alignments.

ACKNOWLEDGMENTS

I thank D. Vertigan and M. Waterman for advice.

REFERENCES

Felsenstein, J. (1978). The number of evolutionary trees. *Syst. Zool.* **27:** 27–33.

Griggs, J. R., Hanlon, P., Odlyzko, A. M., and Waterman, M. S. (1990). On the number of alignments of *k* sequences. *Graphs Combinatorics* **6:** 133–146.

Laquer, H. T. (1981). Asymptotic limits for a two-dimensional recursion. *Stud. Appl. Math.* **64:** 271–277.

Roman, S. (1986). "An Introduction to Discrete Mathematics." Saunders, Philadelphia.

Stanton, R. G., and Cowan, D. D. (1970). Note on a square functional equation. *SIAM Rev.* **12:** 277–279.

Waterman, M. S. (1995). "Maps, Sequences, and Genomes," Chapman and Hall, London.

Wheeler, W. (1996). Optimization alignment: The end of multiple sequence alignment in phylogenetics? *Cladistics* **12:** 1–9.