On the Number of Standard and of Effective Alignments

A. W. M. Dress[1], B. Morgenstern[2], J. Stoye[1,3]

[1] Research Center for Interdisciplinary Studies on Structure Formation (FSPM), University of Bielefeld, Germany
[2] GSF – National Research Center for Environment and Health, Institute of Biomathematics and Biometry, Neuherberg, Germany
[3] Department of Computer Science, University of California at Davis, USA

Introduction

We show how to calculate the number of all possible alignments of N sequences generalizing results of Laquer [1] and Waterman [2] who solved this problem for the special case of $N = 2$ sequences. We consider two notions of sequence alignment: *standard* and *effective* alignments. We present recursive functions to calculate both, the number of standard and the number of effective alignments. We also derive explicit formulae (i) for the number of standard alignments and (ii) for the number of effective alignments of just two sequences.

Terminology

A **standard alignment** of N sequences of length $L_1, \ldots, L_N$ is defined to be an $N \times L$ matrix ($\max(L_1, \ldots, L_N) \leq L \leq \sum_{1 \leq i \leq N} L_i$) whose rows are obtained from the original sequences by insertion of so-called 'blanks' or 'gap characters' – with the additional requirement that no column of the alignment consists exclusively of blanks.

$F(L_1, L_2, \ldots, L_N)$:= the number of standard alignments of N sequences of length $L_1, L_2, \ldots, L_N$.

An **effective alignment** of N sequences of length $L_1, \ldots, L_N$ is a *consistent equivalence relation* defined on the *site space* $\mathcal{S} := \{[i|j] \mid 1 \leq i \leq N, 1 \leq j \leq L_i\}$. This definition avoids a certain redundancy inherent in the standard definition and allows to apply the mathematical theory of sets and relations to investigate the *state space* associated with an alignment problem. (For a more detailed discussion see [3].)

$G(L_1, L_2, \ldots, L_N)$:= the number of effective alignments of N sequences of length $L_1, L_2, \ldots, L_N$.

Summary of Results

First Result A recursive formula for the number of standard alignments:

$$\begin{aligned} F(L_1) &= 1 \\ F(L_1, \ldots, L_{i-1}, 0, L_{i+1}, \ldots, L_N) &= F(L_1, \ldots, L_{i-1}, L_{i+1}, \ldots, L_N) \\ F(L_1, \ldots, L_N) &= \textstyle\sum_{\emptyset \neq V \subseteq \{1,\ldots,N\}} F\left(L_1 - \chi_V(1), \ldots, L_N - \chi_V(N)\right) \end{aligned}$$

where χ_V is the characteristic function

$$\chi_V : \{1, \ldots, N\} \to \{0, 1\} : i \mapsto \begin{cases} 1 & \text{if } i \in V \\ 0 & \text{otherwise} \end{cases}$$

Second Result An explicit formula for the number of standard alignments:

$$F(L_1, \ldots, L_N) = \sum_{L \geq 0} \sum_{x \geq 0} (-1)^x \binom{L}{x} \prod_{i=1}^{N} \binom{L-x}{L_i}$$

Third Result A recursive formula for the number of effective alignments:

$$\begin{aligned} G(L_1) &= 1 \\ G(L_1, \ldots, L_{i-1}, 0, L_{i+1}, \ldots, L_N) &= G(L_1, \ldots, L_{i-1}, L_{i+1}, \ldots, L_N) \\ G(L_1, \ldots, L_N) &= \textstyle\sum_{\emptyset \neq W \subseteq \{1,\ldots,N\}} a(|W|) G\left(L_1 - \chi_W(1), \ldots, L_N - \chi_W(N)\right) \end{aligned}$$

where the numbers a are defined as follows:

$$a(k) := \sum_{\sim} (-1)^{1+\#(\{1,\ldots,k\}/\sim)}$$

and where, for any given $k \in \mathbf{N}_0$, we sum over all equivalence "$\sim$" relations defined on $\{1, \ldots, k\}$, and $\#(\{1, \ldots, k\}/\sim)$ denotes the number of equivalence classes of the equivalence relation "$\sim$".

Fourth Result An explicit formula for the number of effective alignments of two sequences:

$$G(L_1, L_2) = \binom{L_1 + L_2}{L_1} = \binom{L_1 + L_2}{L_2}.$$

Open Question We leave the development of an explicit formula for the number of effective alignments of an arbitrary number of sequences as an open question.

Proofs

Here, we show in full detail only the proof of the second result. The first and the fourth result are quite obvious. The proof of the third result, which – similar to that of the second result – uses Möbius inversion as well as a deeper discussion of the numbers a can be found in [4].

Proof of the Second Result

The idea is to sum over all possible lengths of alignments.

1. Let

   $F(L_1, \ldots, L_N; L)$:= the number of standard alignments of length L of N sequences of length $L_1, L_2, \ldots, L_N$.

   Then

   $$F(L_1, \ldots, L_N) = \sum_{\max(L_1,\ldots,L_N) \leq L \leq L_1 + \ldots + L_N} F(L_1, \ldots, L_N; L).$$

2. For each $X \subseteq \{1, \ldots, L\}$, put

   $f(X, L)$:= the number of alignments of length L with exactly the columns $j \in X$ consisting of blanks only.

   Then

   $$F(L_1, \ldots, L_N; L) = f(\emptyset, L).$$

3. Let

   $f^+(X, L) := f^+(L_1, \ldots, L_N; X, L)$ = the number of alignments of length L with at least the columns $j \in X$ consisting of blanks only.

   Then

   $$f^+(X, L) = \prod_{i=1}^{N} \binom{L - |X|}{L_i}$$

   and

   $$f^+(X, L) = \sum_{X \subseteq Y \subseteq \{1,\ldots,L\}} f(Y, L).$$

4. By Möbius inversion [5], this implies

   $$F(L_1, \ldots, L_N; L) = \sum_{x \geq 0} (-1)^x \binom{L}{x} \prod_{i=1}^{N} \binom{L-x}{L_i}. \qquad (*)$$

   The standard proof for this fact is the following:

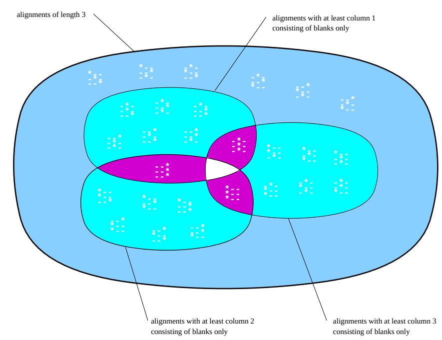
   $$\begin{aligned} \sum_{x \geq 0} (-1)^x \binom{L}{x} \prod_{i=1}^{N} \binom{L-x}{L_i} &= \sum_{X \subseteq \{1,\ldots,L\}} (-1)^{|X|} f^+(X, L) \\ &= \sum_{X \subseteq \{1,\ldots,L\}} (-1)^{|X|} \sum_{X \subseteq Y \subseteq \{1,\ldots,L\}} f(Y, L) \\ &= \sum_{Y \subseteq \{1,\ldots,L\}} f(Y, L) \sum_{X \subseteq Y} (-1)^{|X|} \\ &= f(\emptyset, L) \\ &= F(L_1, \ldots, L_N; L). \end{aligned}$$

5. The final result follows immediately:

   $$\begin{aligned} F(L_1, \ldots, L_N) &= \sum_{\max(L_1,\ldots,L_N) \leq L \leq L_1 + \ldots + L_N} F(L_1, \ldots, L_N; L) \\ &= \sum_{L \geq 0} \sum_{x \geq 0} (-1)^x \binom{L}{x} \prod_{i=1}^{N} \binom{L-x}{L_i}. \end{aligned}$$

□

Informally, one could interpret formula $(*)$ by the Inclusion-Exclusion Principle: To obtain the number of standard alignments of a fixed length L (without blank-only columns), first take the set of all alignments of length L including those with (one or more) columns consisting of blanks only. Since these are more alignments than we want to count, we would like to exclude from these all those alignments which have at least one blank-only column. But we don't have immediate access to their number. Instead, we remove all alignments with at least a blank-only column at position x and then add again the number of alignments which we have excluded more than once, and so on ...
The following figure sketches this principle for alignments of length $L = 3$, given sequences of length $L_1 = 1$, $L_2 = 1$, and $L_3 = 1$.



Discussion

We hope that our work regarding the enumeration of two types of multiple alignments is a first step towards structuring the space of all multiple alignments which will eventually allow to employ well known and highly developed and sophisticated methods from statistical physics to explore the "fitness landscape" defined on that space by various alignment scores, as well as to analyze the various optimization methods designed to actually find their respective (local and/or global) optima.

Acknowledgments

We are grateful to Mike Steel for some useful comments regarding this topic, and we also want to acknowledge that using the World-Wide Web page of [6], `http://www.research.att.com/~njas/sequences/index.html`, proved to be very helpful. J.S. was partly supported by the Graduate College *Strukturbildungsprozesse* (Bielefeld) and by the German Academic Exchange Service (DAAD).

References

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

[2] M. S. Waterman. *Introduction to Computational Biology. Maps, Sequences and Genomes*. Chapman & Hall, London, 1995.

[3] B. Morgenstern, A. W. M. Dress, and T. Werner. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA*, 93(22):12098–12103, 1996.

[4] A. Dress, B. Morgenstern, and J. Stoye. On the number of standard and of effective multiple alignments. *Appl. Math. Lett.*, 1998. To appear.

[5] G.-C. Rota. On the foundations of combinatorial theory I. theory of Möbius functions. *Z. Wahrscheinlichkeitstheorie*, 2:340–368, 1964.

[6] N. J. A. Sloane and S. Plouffe. *The Encyclopedia of Integer Sequences*. Academic Press, San Diego, CA, 1995.