On the number of standard and of effective multiple alignments

A. Dress, B. Morgenstern, J. Stoye

March 30, 1998

Abstract

We study the number of all possible alignments of N sequences, $N \geq 2$, for two distinct alignment concepts proposed in the literature – standard alignments and effective alignments (consistent equivalence relations). Recursion formulae are developed to calculate these numbers. For standard alignments and for effective alignment of just two sequences an explicit formula is also presented. The number of all effective alignments of a given site space is shown to be related to Stirling numbers of second kind.

1 Introduction

Sequence alignment is one of the most important tools for data analysis in molelcular biology. There are different notions of what an alignment is: By standard theory, an alignment of N sequences $s_1, \ldots, s_N$ of length $L_1, \ldots, L_N$ is defined to be an $N \times L$ matrix A with $\max(L_1, \ldots, L_N) \leq L \leq \sum_{1 \leq i \leq N} L_i$ 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 matrix A consists exclusively of blanks (cf. [1]; p. 186).

Recently, Morgenstern *et al.* [2] have proposed a different way of defining alignments (see also [3] and [1], p. 188, for the case of two sequences and [4] for a thorough discussion of this concept for any number of sequences). In their definition, an alignment of the sequences $s_1, \ldots, s_N$ is a *consistent equivalence relation* defined on the so-called *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. To distinguish these alignments from standard alignments, we will refer to them as *effective alignments*.

No matter which definition is preferred, in either case the alignment problem is the problem of finding an *optimal alignment* – according to some well-defined criterion – and the search space for this optimization problem is the set of all possible alignments of a given set of sequences.

Therefore, it seems to be worthwile to study the structure of this space in more detail. In this paper, we show how to calculate the number of all possible alignments of N sequences. We generalize the results of Laquer [5] and Waterman [1] who solved this problem for the special case of $N = 2$ sequences. We derive recursive functions to calculate both, the number of standard alignments and the number of effective alignments. We also present explicit formulae for the number (i) of standard alignments and (ii) of effective alignments of just two sequences.

Although these numerical values themselves are of minor interest to biologists, our study might still be of some use as it sheds light on the structure of the *state space* associated with the alignment problem.

2 The number of standard alignments

Assume that we are given N sequences $s_1, s_2, \ldots, s_N$ of length $L_1, L_2, \ldots, L_N$. Then, clearly, there exist, for any given $L \geq \max(L_1, \ldots, L_N)$, exactly

$$f^+(L) = f^+(L_1, \ldots, L_N; L) := \prod_{i=1}^{N} \binom{L}{L_i}$$

standard alignments of total length L provided we allow columns consisting of blanks, only.

More precisely, given a subset X of $\{1, \ldots, L\}$ of cardinality

$$x \leq L - \max(L_1, \ldots, L_N),$$

there exist

$$f^+(X, L) = f^+(L_1, \ldots, L_N; X, L) := \prod_{i=1}^{N} \binom{L-x}{L_i}$$

such alignments with at least all those columns consisting of blanks only which are indexed by elements $j \in X$.

Consequently, by Möbius inversion [6], the sum

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

coincides with the number $F(L_1, \ldots, L_N; L)$ of all standard alignments of total length L without any column consisting of blanks only.

Remark: The standard proof for this fact runs as follows: for $X \subseteq \{1, \ldots, L\}$ as above, let $f(X, L)$ denote the number of alignments of total length L with exactly those columns consisting of blanks only which are indexed by elements $j \in X$; then, if $x := \#X$, we have

$$f^+(X, L) = \prod_{i=1}^{N} \binom{L-x}{L_i} = \sum_{X \subseteq Y \subseteq \{1, \ldots, L\}} f(Y, L)$$

and hence

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

Clearly, this implies that the number $F(L_1,\ldots,L_N)$ of all standard alignments without any column consisting of blanks only coincides with the double sum

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

where the sum could be taken over all L and x, yet non-zero terms will arise only for $\max(L_1,\ldots,L_N)+x \leq L \leq L_1+\ldots+L_N$.

As any such alignment has a first column involving a well-defined non-empty subset V of $\{1,\ldots,N\}$ of rows without blanks, it is clear that for $L, L_1,\ldots,L_N > 0$, we also have the Pascal-triangle type recursion formulae

$$F(L_1,\ldots,L_N;L) = \sum_{\emptyset\subsetneqq V\subseteq\{1,\ldots,N\}} F(L_1-\chi_V(1),\ldots,L_N-\chi_V(N);L-1)$$

and

$$F(L_1,\ldots,L_N) = \sum_{\emptyset\subsetneqq V\subseteq\{1,\ldots,N\}} F(L_1-\chi_V(1),\ldots,L_N-\chi_V(N))$$

with

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

the characteristic function of $V\subseteq\{1,\ldots,N\}$, as usual. Together with

$$
\begin{aligned}
&F(1;1) = F(1) := 1, \\
&F(1;L) := 0 \text{ for } L > 1,
\end{aligned}
$$

and

$$F(L_1,\ldots,L_N;L) := F(L_1,\ldots,L_{i-1},L_{i+1},\ldots,L_N;L)$$

as well as

$$F(L_1,\ldots,L_N) := F(L_1,\ldots,L_{i-1},L_{i+1},\ldots,L_N)$$

whenever $L_i := 0$ for some $i\in\{1,\ldots,N\}$, this recursion formula can of course also be used to compute the values of $F(L_1,\ldots,L_N;L)$ and $F(L_1,\ldots,L_N)$ in an efficient way.

Remark: Note that a similar argument establishes the recursion formula

$$f^+(L_1,\ldots,L_N;L) = \sum_{V\subseteq\{1,\ldots,N\}} f^+(L_1-\chi_V(1),\ldots,L_N-\chi_V(N);L-1).$$

3 The number of effective alignments

Let us now denote by $G(L_1, \dots, L_N)$ the number of *effective* alignments of the given sequences, that is, of equivalence relations A defined on the set $\mathcal{S} := \{[i|j] \mid 1 \le i \le N, 1 \le j \le L_i\}$ with the property that there exists a partial order $\preceq$ defined on the set $\mathcal{S}/A$ of A-equivalence classes $A(x), A(y), \dots (x, y \in \mathcal{S})$ satisfying the consistency condition

$$(*) \qquad A([i|j]) \preceq A([i|k]) \iff j \le k$$

for all $i \in \{1, \dots, N\}$ and $j, k \in \{1, \dots, L_i\}$. Note that, if any such partial order exists, there exists a unique smallest one which can be defined as the transitive closure of the relation defined by $(*)$ and which will be denoted by "$\preceq_A$".

In case $N = 1$, we clearly have $G(L_1) = 1$; and – just as above – we have

$$G(L_1, \dots, L_N) = G(L_1, \dots, L_{i-1}, L_{i+1}, \dots, L_N)$$

in case $L_i = 0$ for some $i \in \{1, \dots, N\}$. It is also easy to see (cf. [1], p. 188) that, in case $N = 2$, we have

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

because – in view of the identity

$$\begin{aligned} &\sum_{l \ge 0} \binom{L_1 + L_2}{l} x^l = (1+x)^{L_1 + L_2} = (1+x)^{L_1}(1+x)^{L_2} = \\ &= \left(\sum_{l_1 \ge 0} \binom{L_1}{l_1} x^{l_1} \right) \left(\sum_{l_2 \ge 0} \binom{L_2}{l_2} x^{l_2} \right) = \\ &= \sum_{l \ge 0} \left(\sum_{l_1 + l_2 = l} \binom{L_1}{l_1} \binom{L_2}{l_2} \right) x^l \end{aligned}$$

– this number is well known to coincide with

$$\sum_{l_1 + l_2 = L_1} \binom{L_1}{l_1} \binom{L_2}{l_2} = \sum_{l_1 + l_2 = L_1} \binom{L_1}{L_1 - l_1} \binom{L_2}{l_2} = \sum_{k \ge 0} \binom{L_1}{k} \binom{L_2}{k}$$

and because any effective alignment A of two sequences is uniquely determined by the two subsets $K_1 \subseteq \{1, \dots, L_1\}$ and $K_2 \subseteq \{1, \dots, L_2\}$ which are defined by

$$K_1 := \{ j_1 \in \{1, \dots, L_1\} \mid \text{ there exists } j_2 \in \{1, \dots, L_2\} \text{ with } [1|j_1] \overset{A}{\sim} [2|j_2] \}$$

and

$$K_2 := \{ j_2 \in \{1, \dots, L_2\} \mid \text{ there exists } j_1 \in \{1, \dots, L_1\} \text{ with } [2|j_2] \overset{A}{\sim} [1|j_1] \}$$

which can be chosen freely in $\{1,\dots,L_1\}$ and $\{1,\dots,L_2\}$ subject only to the condition that they have to have the same cardinality.

In the general case $N \geq 1$, we can at least derive a Pascal-triangle type recursion formula for $G(L_1,\dots,L_N)$. To this end, consider a *partial partition* $\mathcal{V} = \{V_1,\dots,V_k\}$ of $\{1,\dots,N\}$, that is a non-empty set of non-empty and pairwise disjoint subsets $V_1,\dots,V_k$ of $\{1,\dots,N\}$ and define $G(L_1,\dots,L_N;\mathcal{V})$ to denote the number of effective alignments A for which $\mathcal{V}$ coincides with the set

$$\mathcal{V}(A) := \{V \subseteq \{1,\dots,N\} \bigm| \{[i|1] \bigm| i \in V\} \in \mathcal{S}/A\}.$$

Clearly, $\mathcal{V}(A)$ is non-empty because every A-equivalence class contained in $\mathcal{S}$ which is minimal with respect to the partial order $\preceq_A$ defined by A is necessarily of the form $\{[i|1] \bigm| i \in V\}$ for some non-empty subset $V \subseteq \{1,\dots,N\}$.

So, we have

$$G(L_1,\dots,L_N) = \sum_{\mathcal{V}} G(L_1,\dots,L_N;\mathcal{V}),$$

where the sum is taken over all (non-empty) partial partitions $\mathcal{V}$ of $\{1,\dots,N\}$.

Moreover, if we denote for every such $\mathcal{V}$ by $G^+(L_1,\dots,L_N;\mathcal{V})$ the number of all effective alignments A with $\mathcal{V} \subseteq \mathcal{V}(A)$, we surely have

$$\sum_{\mathcal{V}\subseteq\mathcal{W}} G(L_1,\dots,L_N;\mathcal{W}) = G^+(L_1,\dots,L_N;\mathcal{V}) = G(L_1-\chi_{\underline{\mathcal{V}}}(1),\dots,L_N-\chi_{\underline{\mathcal{V}}}(N))$$

where $\chi_{\underline{\mathcal{V}}}$ denotes the characteristic function of $\underline{\mathcal{V}} := \bigcup_{V\in\mathcal{V}} V$, because that last number just counts the number of effective alignments of the N suffix sequences resulting from our original sequences by eliminating the first entry in each of the sequences s_i with $i \in \underline{\mathcal{V}}$ which is exactly the number of those alignments A of the original sequences with $\mathcal{V} \subseteq \mathcal{V}(A)$.

Consequently, Möbius inversion yields the following recursion formula

$$\begin{aligned}
G(L_1,\dots,L_N;\mathcal{V}) &= \sum_{\mathcal{V}\subseteq\mathcal{W}'} G(L_1,\dots,L_N;\mathcal{W}') \sum_{\mathcal{V}\subseteq\mathcal{W}\subseteq\mathcal{W}'} (-1)^{\#(\mathcal{W}-\mathcal{V})} = \\
&= \sum_{\mathcal{V}\subseteq\mathcal{W}} (-1)^{\#(\mathcal{W}-\mathcal{V})} \sum_{\mathcal{W}\subseteq\mathcal{W}'} G(L_1,\dots,L_N;\mathcal{W}') = \\
&= \sum_{\mathcal{V}\subseteq\mathcal{W}} (-1)^{\#(\mathcal{W}-\mathcal{V})} G^+(L_1,\dots,L_N;\mathcal{W}) = \\
&= \sum_{\mathcal{V}\subseteq\mathcal{W}} (-1)^{\#(\mathcal{W}-\mathcal{V})} G(L_1-\chi_{\underline{\mathcal{W}}}(1),\dots,L_N-\chi_{\underline{\mathcal{W}}}(N))
\end{aligned}$$

which obviously implies the recursion formula

$$
\begin{aligned}
G(L_1,\ldots,L_N) &= \sum_{\emptyset\neq\mathcal{V}} \left(\sum_{\mathcal{V}\subseteq\mathcal{W}} (-1)^{\#(\mathcal{W}-\mathcal{V})} G(L_1-\chi_{\underline{\mathcal{W}}}(1),\ldots,L_N-\chi_{\underline{\mathcal{W}}}(N)) \right) = \\
&= \sum_{\emptyset\neq\mathcal{W}} \left(\sum_{\emptyset\neq\mathcal{V}\subseteq\mathcal{W}} (-1)^{\#(\mathcal{W}-\mathcal{V})} \right) G(L_1-\chi_{\underline{\mathcal{W}}}(1),\ldots,L_N-\chi_{\underline{\mathcal{W}}}(N)) = \\
&= \sum_{\emptyset\neq\mathcal{W}} (-1)^{1+\#\mathcal{W}} G(L_1-\chi_{\underline{\mathcal{W}}}(1),\ldots,L_N-\chi_{\underline{\mathcal{W}}}(N))
\end{aligned}
$$

in view of

$$
\sum_{\emptyset\neq\mathcal{V}\subseteq\mathcal{W}} (-1)^{\#(\mathcal{W}-\mathcal{V})} + (-1)^{\#\mathcal{W}} = \sum_{\mathcal{V}\subseteq\mathcal{W}} (-1)^{\#(\mathcal{W}-\mathcal{V})} = 0.
$$

Moreover, we can rewrite these formulae by introducing the numbers

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

where, for any given $k\in\mathbb{N}_0$, we sum over all equivalence "$\sim$" relations defined on $\{1,\ldots,k\}$. Clearly, we have $a(0)=1, a(1)=-1, a(2)=0, a(3)=1, a(4)=1, a(5)=-2, a(6)=-9, a(7)=-9, a(8)=50$ and so on, as can be read off from the obvious recursion formula

$$
a(k+1) \quad = \quad -\sum_{p=0}^{k} \binom{k}{p} a(k-p).
$$

Remark: The series $a(k)$ also describes the expansion of $\exp(1-e^x)$ and is closely related to the Stirling numbers of second kind σ_k^j [7, 8]: With σ_k^j being the number of equivalence classes with exactly j classes on a set of k distinct elements, we have

$$
a(k) = \sum_{j=1}^{k} (-1)^j \sigma_k^j.
$$

Using these numbers while sorting the above formulae for multiply occuring equal terms, we get

$$
\begin{aligned}
G(L_1,\ldots,L_N;\mathcal{V}) &= \sum_{\mathcal{V}\subseteq\mathcal{W}} (-1)^{\#(\mathcal{W}-\mathcal{V})} G(L_1-\chi_{\underline{\mathcal{W}}}(1),\ldots,L_N-\chi_{\underline{\mathcal{W}}}(N)) = \\
&= \sum_{\underline{\mathcal{V}}\subseteq W} \left(\sum_{\mathcal{V}\subseteq\mathcal{W},\underline{\mathcal{W}}=W} (-1)^{\#(\mathcal{W}-\mathcal{V})} G(L_1-\chi_W(1),\ldots,L_N-\chi_W(N)) \right) = \\
&= \sum_{\underline{\mathcal{V}}\subseteq W\subseteq\{1,\ldots,N\}} \left(\sum_{\sim} (-1)^{\#((W-\underline{\mathcal{V}})/\sim)} \right) G(L_1-\chi_W(1),\ldots,L_N-\chi_W(N)) = \\
&= \sum_{\underline{\mathcal{V}}\subseteq W\subseteq\{1,\ldots,N\}} a(\#(W-\underline{\mathcal{V}})) G(L_1-\chi_W(1),\ldots,L_N-\chi_W(N))
\end{aligned}
$$

as well as

$$G(L_1,\dots,L_N)=\sum_{\emptyset\neq W\subseteq\{1,\dots,N\}} -a(\#W)G(L_1-\chi_W(1),\dots,L_N-\chi_W(N)).$$

In case $N:=2$, this implies

$$\begin{aligned}
G(L_1,L_2;\{\{1\}\}) &= G(L_1-1,L_2)-G(L_1-1,L_2-1),\\
G(L_1,L_2;\{\{2\}\}) &= G(L_1,L_2-1)-G(L_1-1,L_2-1),\\
G(L_1,L_2;\{\{1\},\{2\}\}) &= G(L_1,L_2;\{\{1,2\}\})=G(L_1-1,L_2-1)
\end{aligned}$$

as well as

$$G(L_1,L_2)=G(L_1-1,L_2)+G(L_1,L_2-1)$$

corroborating the result

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

in view of

$$\binom{L_1+L_2}{L_1}=\binom{L_1+L_2-1}{L_1-1}+\binom{L_1+L_2-1}{L_1}=\binom{L_1+L_2-1}{L_1-1}+\binom{L_1+L_2-1}{L_2-1}.$$

In case $N:=3$, we get

$$\begin{aligned}
G(L_1,L_2,L_3;\{\{1\}\}) &= G(L_1-1,L_2,L_3)-G(L_1-1,L_2-1,L_3)-G(L_1-1,L_2,L_3-1)\\
G(L_1,L_2,L_3;\{\{1,2\}\}) &= G(L_1,L_2,L_3;\{\{1\},\{2\}\})\\
&= G(L_1-1,L_2-1,L_3)-G(L_1-1,L_2-1,L_3-1),\\
G(L_1,L_2,L_3;\{\{1,2,3\}\}) &= G(L_1,L_2,L_3;\{\{1,2\},\{3\}\})\\
&= G(L_1,L_2,L_3;\{\{1\},\{2\},\{3\}\})=G(L_1-1,L_2-1,L_3-1)
\end{aligned}$$

as well as

$$\begin{aligned}
G(L_1,L_2,L_3) &= G(L_1-1,L_2,L_3)+G(L_1,L_2-1,L_3)+\\
&\quad +G(L_1,L_2,L_3-1)-G(L_1-1,L_2-1,L_3-1).
\end{aligned}$$

Acknowledgment

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 [7], `http://www.research.att.com/~njas/sequences/index.html`, proved to be very helpful.

References

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

[2] 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).

[3] J.B. Kruskal, An overview of sequence comparison, In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, (Edited by D. Sankoff and J.B. Kruskal), pp. 1-44, Addison-Wesley, Reading (1983).

[4] B. Morgenstern, J. Stoye, and A. Dress, Some theoretical aspects of pairwise and multiple sequence alignment, In preparation.

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

[6] G.-C. Rota, On the foundations of combinatorial theory I. Theory of Möbius functions, *Z. Wahrscheinlichkeitstheorie* **2** 340-368 (1964).

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

[8] V.R.R. Uppuluri and J.A. Carpenter, Numbers generated by the function $\exp(1-e^x)$, *The Fibonacci Quarterly* **7** 437-448 (1969).