Computers in Biology and Medicine 63 (2015) 287–292

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Number of distinct sequence alignments with k-match and match sections

Xiaoqing Liu [a,1], Xiaohua Yang [c,1], Cong Wang [b], Yuhua Yao [b], Qi Dai [b,d,*]

[a] School of Science, Hangzhou Dianzi University, Hangzhou 310018, China
[b] College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
[c] Shangqiu Medical College, Shangqiu 476100, China
[d] Center for Systems Biology, University of Texas at Dallas, Richardson, TX 75080, USA

ARTICLE INFO

Article history:
Received 8 April 2014
Accepted 26 February 2015

Keywords:
Combinatorics
Deletion
Insertion
k-match
Match section
Sequence alignment

ABSTRACT

Background: Recent developments in sequence alignment have led to significant advances in our understanding of the functional, structural or evolutionary relationships among biological sequences. Great efforts have been made to count the total number of sequence alignments, but little attention has been paid to specific alignments associated with conserved patterns.

Methods: We propose a new combinatorial method to count specific alignments. First, we represent a sequence alignment as a system of cells. Using combinatorial techniques and Stirling's formula, we then count the numbers of specific alignments with a *k*-match or match section of size *k*.

Results: We developed three theorems related to different types of specific alignments. We found that the number of the alignments with match sections of at least *k* was less than that of *k*-match sections and the number of specific alignments was significantly lower than the results reported by Covington.

Discussion: The presence of a large number of alignments makes a direct search for the optimal alignment unfeasible for long sequences, whereas our proposed method based on specific alignments decreases the search space by many times. This facilitates the development of a faster algorithm for performing sequence comparisons.

1. Introduction

Since the completion of the Human Genome Project, the development of sequencing technologies has improved the throughput of biological sequences by many times, and thus thousands of biological sequences are generated every day. However, these biological sequences do not increase our understanding of biology, so methods for analyzing these data are increasingly critical as the volume of biological sequence data increases. Sequence comparisons are fundamental operations in bioinformatics and many approaches have been proposed for comparing biological sequences, which can be categorized into two classes: alignment-based methods, where dynamic programming is used to evaluate all possible alignments and select the optimal solution with the highest score [1,2], and alignment-free methods, which measure the similarity between two biological sequences using statistical methods [3–14]. Some alignment-free methods deliver satisfactory performance [3–8], but they are still in the early stages of their development compared with alignment-based methods [9–14].

An alignment of two sequences *a* and *b* should satisfy the following constraints: (1) all of the symbols in the alignment are in the same order as they appear in the sequences *a* and *b*; (2) a symbol from sequence *a* can be aligned with any symbol from sequence *b*; (3) a symbol can also be aligned with a blank "–"; and (4) two blanks are not allowed to align. If a symbol (residue) appears above or below another symbol in an alignment, this indicates that they have a common evolutionary ancestor. We denote them as an identity if they are exactly the same; otherwise, we denote them as a mutation. In this study, we denote both the identities and mutations as matches in an alignment. If a gap appears above or below a symbol, it is assumed that an evolutionary event has occurred, i.e., insertion or deletion.

Many methods can be used to construct an alignment of two biological sequences. To find the optimal alignment, it is necessary to know the total number of distinct alignments that needs to be examined between two biological sequences [15–20]. Laquer studied the number of alignments for two sequences and proposed a general recursion related to the Stanton–Cowan number [21]. Griggs et al.

[*] Corresponding author at: College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China.
E-mail address: daiailiu04@yahoo.com (Q. Dai).

[1] Xiaoqing Liu and Xiaohua Yang contributed equally to this work as co-first authors.

http://dx.doi.org/10.1016/j.compbiomed.2015.02.017
0010-4825/© 2015 Elsevier Ltd. All rights reserved.

studied the number of alignments for t sequences and proved that its exponential growth rate is $(2^{1/t}-1)^{-t}$ [22]. Waterman carefully considered the number of alignments of two sequences and found that they are overestimated [23]. To address this problem, he proposed a new recursion function with some boundary conditions, instead of using the Stanton–Cowan recursion. Using Stirling's formula, he also gave the asymptotic expressions of the recursive function, based on which we know that the direct examination of all alignments is impossible [23,24]. Recently, Covington divided all of the alignments of two sequences into three distinct sets: the largest set, the smallest set and the medium set. Using various distinctiveness criteria, the total number of specific alignments was counted for two sequences [25].

It is well known that evolutionary innovation by mutation is important for adaptation, which ultimately facilitates the survival of all species. In order to highlight the contributions of mutations, their numbers should be a particular focus in an alignment. Moreover, the uninterrupted appearance of matches or a match section usually represents a conservative pattern in biological sequences, which is typically associated with structural and functional domains [26,27]. Therefore, some specific alignments with k-match or match sections of size k should be counted to decrease the size of the search space. In this study, we propose a new combinatorial method for counting specific alignments with k-match or match sections of size k. In order to obtain these counts, we introduce the concepts of cells and a system to represent a sequence alignment. We then use combinatorial techniques to count the specific alignments with k-match or match sections of size k, or at least k.

The remainder of this paper is organized as follows. Section 2 presents the definitions of cells and a system, as well as the lemma and its corollary. Section 3 considers the number of specific alignments for two different sequences with k-match and match sections of size k, or at least k, as well as comparing their differences. Section 4 summarizes the key results of this study.

2. Definition and Lemma

For each alignment, there are two types of symbol: residue and gap. Let 1 indicate the presence of a residue and let 0 denote a gap. First, we define a cell $\begin{bmatrix}x\\y\end{bmatrix}$ as the composition of the sequence alignments. In particular, $\begin{bmatrix}1\\1\end{bmatrix}$ denotes a match cell, $\begin{bmatrix}1\\0\end{bmatrix}$ is a deletion cell and an insertion cell is $\begin{bmatrix}0\\1\end{bmatrix}$. If some cells are arranged in the horizontal direction, a system will be obtained, which corresponds to a sequence alignment. We use this system representation to help count the total number of sequence alignments.

Lemma 1. *If we let $f(n,m)$ = number of alignments of a sequence of length m with a sequence of length n, then*

$$f(n,m)=f(n-1,m)+f(n-1,m-1)+f(n,m-1), \tag{1}$$

with the initial conditions $f(n,0)=f(0,m)=f(0,0)=1$.

Corollary 1. *$F(x,y)$is an ordinary generating function of $f(n,m)$, which is defined as follows:*

$$F(x,y)=\sum_{n\ge0}\sum_{m\ge0} f(n,m)x^ny^m, \tag{2}$$

with the initial conditions $f(n,0)=f(0,m)=f(0,0)=1$, where $(n,m)\ne(0,0)$, and thus

$$F(x,y)=\frac{1}{1-x-y-xy}. \tag{3}$$

Proof. According to Lemma 1

$$\begin{aligned}\sum_{n\ge1}\sum_{m\ge1} f(n,m)x^ny^m &= \sum_{n\ge1}\sum_{m\ge1} f(n-1,m)x^ny^m\\ &\quad+\sum_{n\ge1}\sum_{m\ge1} f(n-1,m-1)x^ny^m\\ &\quad+\sum_{n\ge1}\sum_{m\ge1} f(n,m-1)x^ny^m.\end{aligned} \tag{4}$$

For the left-hand term, we have

$$\begin{aligned}\sum_{n\ge1}\sum_{m\ge1} f(n,m)x^ny^m &= \sum_{n\ge0}\sum_{m\ge0} f(n,m)x^ny^m-\sum_{m\ge0} f(0,m)x^0y^m\\ &\quad-\sum_{n\ge0} f(n,0)x^ny^0+f(0,0)x^0y^0\\ &=F(x,y)-\frac{1}{1-x}-\frac{1}{1-y}+1.\end{aligned} \tag{5}$$

Since

$$\begin{aligned}\sum_{n\ge1}\sum_{m\ge1} f(n-1,m)x^ny^m &= x\left(\sum_{n\ge0}\sum_{m\ge0} f(n,m)x^ny^m-\sum_{n\ge0} f(n,0)x^ny^0\right)\\ &=x\left(F(x,y)-\frac{1}{1-x}\right).\end{aligned} \tag{6}$$

$$\begin{aligned}\sum_{n\ge1}\sum_{m\ge1} f(n-1,m-1)x^ny^m &= xy\left(\sum_{n-1\ge0}\sum_{m-1\ge0} f(n-1,m-1)x^{n-1}y^{m-1}\right)\\ &=xyF(x,y).\end{aligned} \tag{7}$$

$$\begin{aligned}\sum_{n\ge1}\sum_{m\ge1} f(n,m-1)x^ny^m &= y\left(\sum_{n\ge0}\sum_{m\ge0} f(n,m)x^ny^m-\sum_{m\ge0} f(0,m)x^0y^m\right)\\ &=y\left(F(x,y)-\frac{1}{1-y}\right).\end{aligned} \tag{8}$$

By substituting Eq. (5)–(8) into Eq. (4), we obtain

$$F(x,y)=\frac{1}{1-x-y-xy}. \qquad \square \tag{9}$$

3. Counting the number of specific alignments

3.1. Alignments of two different sequences with k-match

Theorem 1. *If we let $f(n,m,k)$ = number of alignments of a sequence of length m with a sequence of length n with k-match, then*

1. $f(n,m,k)=f(n-1,m,k)+f(n-1,m-1,k-1)+f(n,m-1,k)$. (10)

2. $f(n,m,k)=\begin{pmatrix}n+m-k\\k,n-k,m-k\end{pmatrix}$ *where* $0\le k\le \min(n,m)$. (11)

3. $f(n,m,k)\sim\dfrac{1}{2\pi}(r_1+r_2-1)^{k+1/2}\left(1+\dfrac{r_2}{r_1-1}\right)^{(r_1-1)k+1/2}\left(1+\dfrac{r_1}{r_2-1}\right)^{(r_2-1)k+1/2}\dfrac{1}{k(r_1+r_2-1)}$. (12)

   where $r_1=n/k$, $r_2=m/k$, $0<k\le \min(n,m)$.

Proof. (1) Using the system representation, all of the alignments with k-match can be divided into three groups. The alignments of the first group end with a deletion cell $\begin{bmatrix}1\\0\end{bmatrix}$, where the total number is $f(n-1,m,k)$. The alignments in the second group end with a match cell $\begin{bmatrix}1\\1\end{bmatrix}$, and their total number is $f(n-1,m-1,k-1)$. For the last group, the alignments end with an insertion cell $\begin{bmatrix}0\\1\end{bmatrix}$, and their total number is $f(n,m-1,k)$. Thus,

$$f(n,m,k)=f(n-1,m,k)+f(n-1,m-1,k-1)+f(n,m-1,k).$$