DNA Sequence

ISSN: 1042-5179 (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/imdn19

An Exact Formula for the Number of Alignments Between Two DNA Sequences

Angela Torres, Alberto Cabada & Juan J. Nieto

To cite this article: Angela Torres, Alberto Cabada & Juan J. Nieto (2003) An Exact Formula for the Number of Alignments Between Two DNA Sequences, DNA Sequence, 14:6, 427-430, DOI: 10.1080/10425170310001617894

To link to this article: https://doi.org/10.1080/10425170310001617894

Published online: 11 Jul 2009.

DNA Sequence, December 2003 Vol. 14 (6), pp. 427–430

Taylor & Francis
healthsciences

Short Communication

An Exact Formula for the Number of Alignments Between Two DNA Sequences

ANGELA TORRES[a], ALBERTO CABADA[b] and JUAN J. NIETO[b,*]

[a]*Departamento de Psiquiatría, Radiología y Salud Pública, Facultad de Medicina, Universidad de Santiago de Compostela, Santiago de Compostela, Spain;* [b]*Departamento de Análisis Matemático, Facultad de Matemáticas, Universidad de Santiago de Compostela, Santiago de Compostela, Spain*

(Received 9 June 2003)

Given two DNA sequences, one is usually interested in measuring their similarity. For this purpose we have to consider different alignments of both the sequences. The number of alignments grows rapidly with the length of the sequences. In this short communication, we give an exact formula for the number of possible alignments using the theory of difference equations.

Keywords: DNA sequence; Alignment; Difference equation; Exact formula

INTRODUCTION

Mathematics, statistics and computer science play a relevant role in the study of genetic sequences and, generally, in molecular biology, bioinformatics and computational biology (Paun *et al.*, 1998; Tang, 2000; Jamshidi *et al.*, 2001; Lange, 2002; Morgenstern, 2002; Percus, 2002; Nieto *et al.*, 2003; Torres and Nieto, 2003).

The comparison of genomic sequences (Forster *et al.*, 1999; Gusev *et al.*, 1999; Li *et al.*, 2001; Liben-Nowell, 2001; Jiang *et al.*, 2002) is performed using sequence alignment as a basic tool. It is obvious that the number of possible alignments grows rapidly. Here, we present an exact formula for the number of alignments using the theory of difference equations and a computer program to compute that number, and discuss some of its implications. Our method gives an alternate way for computing the number of possible gapped alignments. For example, for two sequences of length 8 and 16, respectively, the number of alignments is $\approx 40 \times 10^6$. For two sequences of length greater than 107, such a number is greater than 10^{80}.

ALIGNMENTS

The information archive in each organism is the genetic material, DNA or, in some viruses, RNA. DNA molecules are long chain molecules containing a message in a four-letter alphabet: A (adenine), G (guanine), C (cytosine) and T (thymine). In the RNA alphabet, T is replaced by U (uracil). The genetic information is encoded digitally as strings over this four-letter alphabet. A DNA sequence is just a string

$$x = (x_1, x_2, \ldots, x_n)$$

where x_i, $i = 1, 2, \ldots, n$ is one of the four nucleotides A, G, C, T. We say that the sequence x is of length n.

Given another sequence

$$y = (y_1, y_2, \ldots, y_m),$$

we wish to compare both sequences and measure their similarity and determine their residue–residue correspondences. To compare the nucleotides or amino acids that appear at corresponding positions in two sequences, we must first assign those correspondences and a sequence alignment is just the identification of residue–residue correspondence (Lesk, 2002). Any assignment of correspondences that preserves the order of the residues within the

*Corresponding author. Tel.: +34-981-563-100. Fax: +34-981-597-054. E-mail: amnieto@usc.es

ISSN 1042-5179 print/ISSN 1029-2365 online © 2003 Taylor & Francis Ltd
DOI: 10.1080/1042517031000161789 4

sequences is an alignment and we allow the introduction of gaps. Thus, an alignment of sequences x and y is an arrangement of $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ and gaps in two rows, keeping the ordering on the x_i and on the y_i, but not having any gap directly over another gap. This restriction implies that there are only a finite number of possible alignments.

For example, given the sequences $x = \mathrm{CGT}$ and $y = \mathrm{ACTT}$, an alignment is

–	C	G	T	–
A	C	–	T	T

Sequence alignments is one of the basic tools of molecular biology, computational biology and bioinformatics (Altschul *et al.*, 2001; Lesk, 2002; Sadovsky, 2003; Torres and Nieto, 2003).

The number of alignments $f(n,m)$ between two given sequences $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_m)$ is given by the following recurrence relation (see Lange, 2002),

$$f(n,m) = f(n-1,m) + f(n,m-1) + f(n-1,m-1), \quad (1)$$

together with the initial conditions

$$f(0,m) = f(n,0) = 1, \quad n,m \in \{1,2,3,\ldots\}. \quad (2)$$

The initial conditions hold from the trivial fact that the only alignment is given by insert n blanks in the string with none elements. The general recurrence formula follows from the fact that every pair of modified sequences ends with one of the three pairs

$$\begin{pmatrix} x_n \\ - \end{pmatrix}, \quad \begin{pmatrix} - \\ y_m \end{pmatrix} \text{ or } \begin{pmatrix} x_n \\ y_m \end{pmatrix},$$

and the remainder of the alignment to the left of one of these pairs constitutes an alignment between two shorter sequences.

EXACT FORMULA

It is obvious that Eqs. (1) and (2) have a unique solution since Eq. (1) is an explicit recurrence relation with given initial conditions (2). To solve such problem we shall use some results from the theory of difference equations (Kelley and Peterson, 2001). One of the advantages of our method is that it scales better than direct evaluation of Eq. (1). Denote for $n \in \{1,2,\ldots\}$, $k \in \{0,1,\ldots\}$, the "falling factorial power" as $n^{\underline{0}} = 1$ and, if $k \geq 1$,

$$n^{\underline{k}} = n(n-1)\cdots(n-k+1). \quad (3)$$

From the definition, one can verify the following property

$$(n+1)^{\underline{k}} - n^{\underline{k}} = k n^{\underline{k-1}}, \quad k \geq 0. \quad (4)$$

First, we assume that $n \geq m$. We look for a solution given by the following expression

$$f(n,m) = \sum_{i=0}^{m+1} A_i(m) n^{\underline{i}}, \quad (5)$$

where

$$A_0(m) \equiv 1, \quad \text{for all } m \in \{0,\ldots,n\}. \quad (6)$$

Note that if the solution f is given by this expression then

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

$$= \sum_{i=0}^{m+1} A_i(m)((n+1)^{\underline{i}} + n^{\underline{i}}).$$

Using property (4) we deduce that this last expression is equal to

$$\sum_{i=1}^{m+1} (2A_{i-1}(m) + iA_i(m)) n^{\underline{i-1}} + 2A_{m+1}(m) n^{\underline{m+1}}$$

As consequence of these equalities, from property (4) and the initial conditions (2), we conclude that

$$f(n,m+1) = 1 + \sum_{i=1}^{m+1} (2A_{i-1}(m) + iA_i(m)) \frac{n^{\underline{i}}}{i}$$

$$+ 2A_{m+1}(m) \frac{n^{\underline{m+2}}}{m+2}.$$

From this last expression and the equality (5) applied at the point $(n, m+1)$, we have that the coefficients $A_i(m)$ satisfy the following recurrence system

$$A_i(m+1) = A_i(m) + \frac{2}{i} A_{i-1}(m),$$

$$i \in \{1,2,\ldots,m+1\}, \quad (7)$$

and

$$A_{m+2}(m+1) = \frac{2}{m+2} A_{m+1}(m). \quad (8)$$

Now, taking into account that $1 = f(n,0) = 1 + A_1(0)n$, we arrive at $A_1(0) = 0$, and as a consequence of Eq. (8), we have that the following property holds

$$A_i(i-1) = 0 \quad \text{for all } i \geq 1. \quad (9)$$

Now, we rewrite Eq. (7) with initial conditions (9) as

$$A_i(m+1) = A_i(m) + A_{i-1}(m), \ m \geq i-1, \ \ A_i(i-1) = 0.$$

In the sequel, we verify that the unique solution of this problem is given by

$$A_i(m) = \frac{2^i}{(i!)^2} m^{\underline{i}} \quad \text{for all } i \in \{0,1,\ldots,m\}.$$

For $i = 1$, since $A_0(m) = 1$ for all $m \geq 1$ we have that $A_1(m) = 2m \equiv 2m^{\underline{1}}$.

Assume that $i \geq 2$. The initial condition follows trivially from the fact that $(i-1)^{\underline{i}} = 0$. Equation (7) follows directly from property (4).

When $n \leq m$, we look for a solution that satisfies the following expression

$$f(n,m) = \sum_{i=0}^{n+1} A_i(n) m^{\underline{i}},$$

and conclude that $f(n,m) = f(m,n)$.

Thus, using that if $0 \leq k \leq n$, then

$$n^{\underline{k}} = k! \binom{n}{k},$$

as a consequence of the development done in this section, we arrive at the following result.

Conclusion: The unique solution of problems (1) and (2) is given by the following expression

$$f(n,m) = \sum_{k=0}^{\min\{n,m\}} 2^k \binom{m}{k} \binom{n}{k}. \qquad (10)$$

We note that the formula allows for the situation where no residues are aligned between sequences, i.e. all the residues of one sequence are in the gaps of the other sequence. Of course, this is a non-biological result and there are a huge number of such alignments. In a forthcoming paper we will present the formula for the number of alignments with no overlapping residues, the number of optimal alignments (with respect to an optimality criterion), and the expression for the number of multiple alignments. Also, we will discuss explicitly some scaling questions in the context of biological sequences.

CONSEQUENCES

Equation (10) permeates us to compute directly the value of the alignments between two strings of DNA. In the sequel we give some numerical values for such combinations

$$f(2,1) = f(1,2) = 5,$$

$$f(4,2) = f(2,4) = 41,$$

$$f(8,4) = f(4,8) = 3649,$$

$$f(16,8) = f(8,16) = 39490049.$$

Hence, we can see in this simple example that the number of alignments grows very fast. Equation (10) shows that the double sequence $\{f(n,m)\}$ is monotone nondecreasing in both variables, i.e. $f(n,m) \leq f(n,m+1)$ and $f(n,m) \leq f(n+1,m)$, for all $n, m \in N$. In particular, $f(n,m) = f(m,n) \geq f(n,n)$ for all $n \leq m$ and $f(n,m) = f(m,n) \leq f(n,n)$ whenever $n \geq m$.

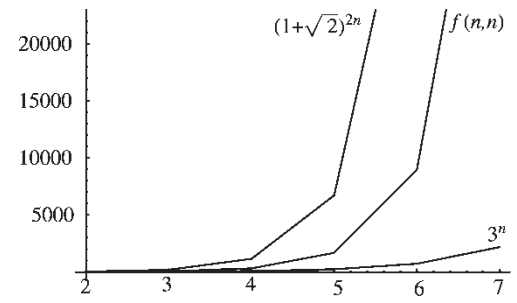

FIGURE 1 Upper and lower estimation of the growth of the number of alignments $f(n,n)$.

From this fact, the behaviour of function f at the diagonal terms (n,n) plays an important role in the study of the considered functions. Clearly, $\{f(n,n)\}$ is a nondecreasing monotone. And it grows very fast, as we can note in the following facts

$$f(10,10) = 8097453,$$

$$f(50,50) \approx 15 \times 10^{36},$$

$$f(100,100) \approx 2 \times 10^{74},$$

Special mention of the inequalities

$$f(106,106) < 10^{80} < f(107,107).$$

This last estimation is relevant since there are only 10^{80} protons in our universe (see Nowak and May, 2000, p. 84). A typical sequence contains about 200–500 base pairs (Lesk, 2002). Proteins are about 200–400 amino acids long, and hence they are 600–1200 letters long.

Using the Newton's binomial formula and classical inequalities, we arrive at the following estimations from above and below on the growth of $f(n,n)$:

$$\sum_{k=0}^{n} 2^k \binom{n}{k} = 3^n \leq f(n,n)$$

$$\leq \left(\sum_{k=0}^{n} \sqrt{2}^k \binom{n}{k}\right)\left(\sum_{k=0}^{n} \sqrt{2}^k \binom{n}{k}\right)$$

$$= (1+\sqrt{2})^{2n}.$$

These estimations are shown in Fig. 1 when both sequences are of the same length n.

Acknowledgements

Second and third author's research partially supported by Ministerio de Ciencia y Tecnología and FEDER, project BFM2001-3884-C02-01, and by Xunta de Galicia and FEDER, project PGIDIT02PXIC20703PN. The authors thank the anonymous reviewers for their useful comments.

References

Altschul, S.F., Bundschuh, R., Olsen, R. and Hwa, T. (2001) "The estimation of statistical parameters for local alignment score distributions", *Nucleic Acid Research* **29**, 351–361.

Forster, M., Heath, A. and Afzal, M. (1999) "Application of distance geometry to 3D visualization of sequence relationships", *Bioinformatics* **15**, 89–90.

Gusev, V.D., Nemytikova, L.A. and Chuzhanova, N.A. (1999) "On the complexity measures of genetic sequences", *Bioinformatics* **15**, 994–999.

Jamshidi, N., Edwards, J.S., Fahland, T., Church, G.M. and Palsson, B.O. (2001) "Dynamic simulation of the human red blood cell metabolic network", *Bioinformatics* **17**, 286–287.

Jiang, T., Lin, G., Ma, B. and Zhang, K. (2002) "A general edit distance between RNA structures", *Journal of Computational Biology* **9**, 371–388.

Kelley, W.G. and Peterson, A.C. (2001) *Difference Equations. An Introduction with Applications* (Academic Press, San Diego).

Lange, K. (2002) *Mathematical and Statistical Methods for Genetic Analysis* (Springer-Verlag, New York).

Lesk, A.M. (2002) *Introduction to Bioinformatics* (Oxford University Press, Oxford).

Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P. and Zhang, H. (2001) "An information-based sequence distance and its application to whole mitochondrial phylogeny", *Bioinformatics* **17**, 149–154.

Liben-Nowell, D. (2001) "On the structure of syntenic distance", *Journal of Computational Biology* **8**, 53–67.

Morgenstern, B. (2002) "A simple and space-efficient fragment-chaining algorithm for alignment of DNA and protein sequences", *Applied Mathematics Letters* **15**, 11–16.

Nieto, J.J., Torres, A. and Vázquez-Trasande, M.M. (2003) "A metric space to study differences between polynucleotides", *Applied Mathematics Letters*, in press.

Nowak, M.A. and May, R.M. (2000) *Virus Dynamics* (Oxford University Press, Oxford).

Paun, Gh., Rozenberg, G. and Salomaa, A. (1998) *DNA Computing: New Computing Paradigms* (Springer-Verlag, Berlin).

Percus, J. (2002) *Mathematics of Genome Analysis* (Cambridge University Press, Cambridge).

Sadovsky, M.G. (2003) "The method to compare nucleotide sequences based on the minimum entropy principle", *Bulletin of Mathematical Biology* **65**, 309–322.

Tang, B. (2000) "Evaluation of some DNA cloning strategies", *Computers and Mathematics with Applications* **39**, 43–48.

Torres, A. and Nieto, J.J. (2003) "The fuzzy polynucleotide space: basic properties", *Bioinformatics* **19**, 587–592.