Andrade *et al. BMC Bioinformatics* 2014, **15**:94
http://www.biomedcentral.com/1471-2105/15/94

BMC Bioinformatics

RESEARCH ARTICLE Open Access

The number of reduced alignments between two DNA sequences

Helena Andrade[1], Iván Area[2], Juan J Nieto[1,3*] and Ángela Torres[4]

Abstract

Background: In this study we consider DNA sequences as mathematical strings. Total and reduced alignments between two DNA sequences have been considered in the literature to measure their similarity. Results for explicit representations of some alignments have been already obtained.

Results: We present exact, explicit and computable formulas for the number of different possible alignments between two DNA sequences and a new formula for a class of reduced alignments.

Conclusions: A unified approach for a wide class of alignments between two DNA sequences has been provided. The formula is computable and, if complemented by software development, will provide a deeper insight into the theory of sequence alignment and give rise to new comparison methods.

AMS Subject Classification: Primary 92B05, 33C20, secondary 39A14, 65Q30

Keywords: DNA sequence, Alignment, Difference equation

Background

Let us consider a DNA sequence as a mathematical string

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

where $x_i \in \{A, G, C, T\}$ is one of the four nucleotides, $i = 1, 2, \ldots, n$, i.e. A denotes adenine, C cytosine, G guanine and T thymine. In these conditions, the sequence x is of length n.

Our main goal is to compare the sequence x with another DNA sequence

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

to measure the similarity between both strings and also to determine their residue-residue correspondences.

Sequence comparison and alignment is a central and crucial tool in molecular biology. For example, Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid) [1].

For some recent developments and directions we refer the reader to [2-7] and [8] for a general review of different alignments methods.

To align the sequences CGT and $ACTT$, one can use EMBOSS Needle for nucleotide sequence [9] that creates an optimal global alignment of the two sequences using the Needleman-Wunsch algorithm to get

```
EMBOSS-001  1  -  C  G  T  3
                  |  ·  |
EMBOSS-001  1  A  C  T  T  4
```

Following Lesk [10], in order to compare the amino acids appearing at their corresponding positions in two sequences, theirs correspondences must be assigned and a sequence alignment is the identification of residue-residue correspondence. For some references on sequence alignment we refer the reader to [10-16].

To compare two sequences, there exist mainly three different possibilities leading to three different numbers of total alignments [10,11,13]:

*Correspondence: juanjose.nieto.roig@usc.es
[1]Departamento de Análise Matemática, Facultade de Matemáticas, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
[3]Faculty of Science, King Abdulaziz University, P.O. Box 80203, 21589 Jeddah, Saudi Arabia
Full list of author information is available at the end of the article

1. The total number of alignments denoted by $f(n,m)$ that was solved in [13].
2. A gap in a sequence is followed by another gap in the other sequence as in Alignments 1 and 2 for the sequences $x = CGT$ and $y = ACTT$ (see Tables 1 and 2 below)
   Considering the two alignments as equivalents to the Alignment 3 (see Table 3) without gap in those positions, we have the number of reduced alignments denoted by $h(n,m)$, and obviously $h(n,m) < f(n,m)$. This case has been solved in [11], and we give here another representation in terms of hypergeometric series.
3. In the interesting case that the alignments 1 and 2 are equivalent, but different from alignment 3 we have a number or reduced alignments $g(n,m)$ where $h(n,m) < g(n,m) < f(n,m)$. This last case is new and we present an explicit formula for g.

Results and discussion

Number of $f(x,y)$ alignments

The total number of alignments $f(x,y)$ satisfies the following recurrence relation [13]

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

with initial conditions $f(n,0) = f(0,m) = 1$ for $n,m = 1,2,3,\ldots$. The solution of the above partial difference equation is given by

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

(see formula (10) in [13]) and the generating function [17,18] is

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

Therefore the coefficients $f(n,m)$ in the expansion

$$F(x,y) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} f(n,m)x^n y^m$$

are given in terms of a hypergeometric series by

$$f(n,m) = {}_2F_1(-m,-n;1;2).$$

This relation seems to be new in this form. Here, the generalized hypergeometric series is defined as (see e.g. [19, Chapter 16])

$${}_pF_q\left(a_1,\ldots,a_p;b_1,\ldots,b_q;d\right) = \sum_{i=0}^{\infty} \frac{(a_1)_k (a_2)_k \cdots (a_p)_k}{k!\,(b_1)_k (b_2)_k \cdots (b_q)_k} d^k,$$

Table 1 Alignment 1

C	G	–	T	–
A	–	C	T	T

Table 2 Alignment 2

C	–	G	T	–
A	C	–	T	T

and $(A)_k = A(A+1)\cdots(A+n-1)$, with $(A)_0 = 1$, denotes the Pochhammer's symbol. It is assumed that $b_j \neq -k$ in order to avoid singularities in the denominators. If one of the parameters a_j equals to a negative integer, then the sum becomes a terminating series.

Number of $h(x,y)$ alignments

In this case, the recurrence relation for the $h(n,m)$ coefficients is [11]

$$h(n,m) = h(n-1,m) + h(n,m-1) - h(n-2,m-2), \quad n,m \geq 2,$$

with initial conditions $h(n,0) = h(0,m) = 1$. Therefore, the generating function [17,18] is

$$H(x,y) = \frac{1-xy}{x^2y^2 - x - y + 1},$$

and the coefficients in the expansion

$$H(x,y) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} h(n,m)x^n y^m$$

are given by

$$h(n,m) = \sum_{i=0}^{A} \frac{(-1)^i(-3i+m+n)!}{i!\,(m-2i)!\,(n-2i)!} - \sum_{i=0}^{B} \frac{(-1)^i(-3i+m+n-2)!}{i!\,(-2i+m-1)!\,(-2i+n-1)!},$$

where

$$A = \min\left\{\left[\frac{n}{2}\right],\left[\frac{m}{2}\right]\right\},$$
$$B = \min\left\{\left[\frac{n-1}{2}\right],\left[\frac{m-1}{2}\right]\right\}.$$

The above coefficients can be written in terms of (terminating) hypergeometric series as

$$\frac{(m+n)!}{m!\,n!}\,{}_4F_3\left(\left.\begin{matrix}\frac{1-m}{2}, -\frac{m}{2}, \frac{1-n}{2}, -\frac{n}{2} \\ \frac{-m-n}{3}, \frac{-m-n+1}{3}, \frac{-m-n+2}{3}\end{matrix}\right| \frac{16}{27}\right) - \frac{(m+n-2)!}{(m-1)!\,(n-1)!}\,{}_4F_3\left(\left.\begin{matrix}\frac{1-m}{2}, 1-\frac{m}{2}, \frac{1-n}{2}, 1-\frac{n}{2} \\ \frac{-m-n+2}{3}, \frac{-m-n+3}{3}, \frac{-m-n+4}{3}\end{matrix}\right| \frac{16}{27}\right).$$

Table 3 Alignment 3

C	G	T	–
A	C	T	T

Number of $g(x, y)$ alignments

As indicated before, the main aim of this paper is to give an explicit representation in this case. The recurrence relation for the $g(n, m)$ coefficients is [11]

$$g(n,m) = g(n-1,m-1) + g(n-1,m) + g(n,m-1) - 2g(n-2,m-2), \quad n,m \geq 2,$$

with initial conditions $g(n,0) = g(m,0) = 1$. Thus, the generating function [17,18] is

$$G(x,y) = \frac{1-xy}{2x^2y^2 - xy - x - y + 1}. \tag{1}$$

Theorem 1. *The coefficients $\alpha_{n,m}$ in the expansion*

$$G(x,y) = \frac{1-xy}{2x^2y^2 - xy - x - y + 1} = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \alpha_{n,m} x^n y^m \tag{2}$$

are explicitly given by

$$\alpha_{n,m} = \left(\sum_{i=U(n,m)}^{n+m} \sum_{j=A(i,n,m)}^{B(i,n,m)} \beta_{i,j,n,m} \right) - \left(\sum_{i=U(n,m)-1}^{n+m-2} \sum_{j=C(i,n,m)}^{D(i,n,m)} \gamma_{i,j,n,m} \right), \tag{3}$$

where

$$\beta_{i,j,n,m} = \frac{(-1)^{i-j} 2^{i-j}\, i!}{(i-j)!\ (2i-j-m)!\ (2i-j-n)!\ (3j-4i+m+n)!}, \tag{4}$$

$$\gamma_{i,j,n,m} = \frac{(-1)^{i-j} 2^{i-j}\, i!}{(i-j)!\,(2i-j-m+1)!\ (2i-j-n+1)!\ (3j-4i+m+n-2)!}, \tag{5}$$

$$A(i,n,m) = \max\left\{0, \left\lceil \frac{4i-m-n}{3} \right\rceil \right\}, \tag{6}$$

$$B(i,n,m) = \min\left\{i, 2i-m, 2i-n, \left\lceil \frac{4i-n-m}{2} \right\rceil \right\}, \tag{7}$$

$$C(i,n,m) = \max\left\{0, \left\lceil \frac{4i-m-n-2}{3} \right\rceil \right\}, \tag{8}$$

$$D(i,n,m) = \min\left\{i, 2i-m+1, 2i-n+1, \left\lceil \frac{4i-n-m+2}{2} \right\rceil \right\}, \tag{9}$$

$$U(n,m) = \begin{cases} m - [n/2], & n \leq m, \\ [(m+1)/2] + n - m, & n \geq m, \end{cases} \tag{10}$$

and $[x]$ denotes the integer part of x.

Proof. If we expand,

$$G(x,y) = (1-xy)\sum_{i=0}^{\infty} \left(x + y + xy - 2x^2y^2\right)^i = (1-xy) \times \sum_{i=0}^{\infty} \left(\sum_{j=0}^{i} \left(\sum_{k=0}^{j} \left(\sum_{s=0}^{k} (-1)^{i-j} 2^{i-j} \binom{i}{j}\binom{j}{k} \binom{k}{s} y^{2i-j-s} x^{2i-j-k+s} \right) \right) \right), \tag{11}$$

we have two summands to be computed, namely

$$\sum_{i=0}^{\infty} \left(\sum_{j=0}^{i} \left(\sum_{k=0}^{j} \left(\sum_{s=0}^{k} (-1)^{i-j} 2^{i-j} \binom{i}{j}\binom{j}{k} \binom{k}{s} y^{2i-j-s} x^{2i-j-k+s} \right) \right) \right) \tag{12}$$

$$- xy \sum_{i=0}^{\infty} \left(\sum_{j=0}^{i} \left(\sum_{k=0}^{j} \left(\sum_{s=0}^{k} (-1)^{i-j} 2^{i-j} \binom{i}{j}\binom{j}{k} \binom{k}{s} y^{2i-j-s} x^{2i-j-k+s} \right) \right) \right). \tag{13}$$

In order to compute the first sum (12) let us introduce

$$m = 2i - j - s, \quad n = 2i - j - k + s. \tag{14}$$

Therefore, the summation to be done reads as

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \left(\sum_{i=U}^{V}\sum_{j=A}^{B} (-1)^{i-j} 2^{i-j} \binom{i}{j} \binom{j}{4i-2j-m-n} \binom{4i-2j-m-n}{2i-j-m} \right) x^n y^m$$

where U, V, A and B must be computed in terms of the initial indices.

The product of binomials can be simplified to

$$\frac{i!}{(i-j)!\ (2i-j-m)!\ (2i-j-n)!\ (3j-4i+m+n)!}.$$

Thus,

$$i \geq 0, \quad j \geq 0, \quad 4i-2j-m-n \geq 0, \quad 4i-2j-m-n \geq 0, \quad 2i-j-m \geq 0, \quad i-j \geq 0, \quad 2i-j-m \geq 0, \quad 2i-j-n \geq 0, \quad 3j-4i+m+n \geq 0,$$

and then

$$A(i,n,m) = A = \max\left\{0, \left[\frac{4i-m-n}{3}\right]\right\} \le j \le \min\left\{i, 2i-m, 2i-n, \left[\frac{4i-n-m}{2}\right]\right\} = B(i,n,m) = B.$$

Finally, the summation reads as

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\left(\sum_{i=U(n,m)}^{n+m}\sum_{j=A}^{B}\frac{(-1)^{i-j}2^{i-j}\,i!}{(i-j)!\;(2i-j-m)!\;(2i-j-n)!\;(3j-4i+m+n)!}\right)x^n y^m,$$

where

$$U(n,m) = \begin{cases} m - [n/2], & n \le m, \\ [(m+1)/2] + n - m, & n \ge m. \end{cases}$$

A similar work with the second summand (13) leads to the final result. □

Some numerical values are $g(10,10) = 2003204$, $g(50,50) = 2.71972 \times 10^{34}$, $g(100,100) = 7.55997 \times 10^{69}$, and we note that $g(n,n) > 10^{80}$ for $n \ge 115$. This last inequality is relevant since 10^{80} is an estimation of the number of protons of our universe [13].

Conclusions

A unified approach for a wide class of alignments between two DNA sequences has been provided. We conclude also that our approach gives an explicit formula filling a gap in the theory of sequence alignment. The formula is computable and, if complemented by software development, will provide a deeper insight into the theory of sequence alignment and give rise to new comparison methods. It may be used also, in the future, to get explicit formulas and compute the number of total, reduced, and effective alignments for multiple sequences.

Methods

We have performed a number of numerical computations to compare our formulae and Mathematica® [20] command Coefficient for the series expansion of (1), on a MacBook Pro featuring a 45 nm "Penryn" 2.66 GHz Intel "Core 2 Duo" processor (P8800), with two independent processor "cores" on a single silicon chip, 8 GB of 1066 MHz DDR3 SDRAM (PC3-8500). We would like to mention that our approach is amazingly fast, since e.g. $g(100,100)$ is computed by using Mathematica® in 0.125165 seconds by using the new formulas presented in this paper, while the use of Mathematica® command Coefficient needs 99.167659 seconds.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Each of the authors HA, IA, JJN and AT, contributed to each part of this study equally and read and approved the final version of the manuscript.

Acknowledgements
The authors are grateful to Prof. Marko Petkovšek for helpful comments. The work of I. Area has been partially supported by the Ministerio de Economía y Competitividad of Spain under grant MTM2012–38794–C02–01, co-financed by the European Community fund FEDER. J.J. Nieto also acknowledges partial financial support by the Ministerio de Economía y Competitividad of Spain under grant MTM2010–15314, co-financed by the European Community fund FEDER.

Author details
[1]Departamento de Análise Matemática, Facultade de Matemáticas, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain. [2]Departamento de Matemática Aplicada II, E.E. Telecomunicación, Universidade de Vigo, 36310 Vigo, Spain. [3]Faculty of Science, King Abdulaziz University, P.O. Box 80203, 21589 Jeddah, Saudi Arabia. [4]Departamento de Psiquiatría, Radioloxía e Saúde Pública, Facultade de Medicina, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain.

Received: 10 January 2014 Accepted: 19 March 2014
Published: 1 April 2014

References

1. The European Bioinformatics Institute: **Pairwise Sequence Alignment.** http://www.ebi.ac.uk/Tools/psa/.
2. Orobitg M, Lladós J, Guirado F, Cores F, Notredame C: **Scalability and accuracy improvements of consistency-based multiple sequence alignment tools.** In *EuroMPI*. Edited by Dongarra J, Blas JG, Carretero J. New York, USA: ACM International Conference Proceeding Series; 2013:259–264.
3. Orobitg M, Cores F, Guirado F, Roig C, Notredame C: **Improving multiple sequence alignment biological accuracy through genetic algorithms.** *J Supercomput* 2013, **65**(3):1076–1088.
4. Montañola A, Roig C, Guirado F, Hernández P, Notredame C: **Performance analysis of computational approaches to solve multiple sequence alignment.** *J Supercomput* 2013, **64**(1):69–78.
5. Zhong C, Zhang S: **Efficient alignment of rna secondary structures using sparse dynamic programming.** *BMC Bioinformatics* 2013, **14:**269.
6. Veeneman BA, Iyer MK, Chinnaiyan AM: **Oculus: faster sequence alignment by streaming read compression.** *BMC Bioinformatics* 2012, **13:**297.
7. Chaisson M, Tesler G: **Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): theory and application.** *BMC Bioinformatics* 2012, **13:**238.
8. Löytynoja A: **Alignment methods: Strategies, challenges, benchmarking, and comparative overview.** In *Evolutionary Genomics. Methods in Molecular Biology. Volume 855*. Edited by Anisimova M. New York, USA: Humana Press; 2012:203–235.
9. The European Bioinformatics Institute: **Pairwise Sequence Alignment (Nucleotide).** http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html.
10. Lesk AM: *Introduction to Bioinformatics*. Oxford, UK: Oxford University Press; 2002.
11. Andrade H: **Análise matemática dalgunhos problemas no estudo de secuencias biolóxicas.** PhD thesis, Universidade de Santiago de Compostela, Departamento de Análise Matemática (2013).
12. Bai F, Zhang J, Zheng J: **Similarity analysis of DNA sequences based on the EMD method.** *Appl Math Lett* 2011, **24**(2):232–237.
13. Cabada A, Nieto JJ, Torres A: **An exact formula for the number of aligments between two DNA sequences.** *DNA Sequence (continued as Mitochondrial DNA)* 2003, **14:**427–430.
14. Eger S: **Sequence alignment with arbitrary steps and further generalizations, with applications to alignments in linguistics.** *Inform Sci* 2013, **237:**287–304.

15. Morgenstern B: **A simple and space-efficient fragment-chaining algorithm for alignment of DNA and protein sequences.** *Appl Math Lett* 2002, **15**(1):11–16.
16. Zhang J, Wang R, Bai F, Zheng J: **A quasi-MQ EMD method for similarity analysis of DNA sequences.** *Appl Math Lett* 2011, **24**(12):2052–2058.
17. Srivastava HM, Manocha HL: *A Treatise on Generating Functions. Ellis Horwood Series: Mathematics and its Applications*. Chichester: Ellis Horwood Ltd.; 1984.
18. Wilf HS: *Generatingfunctionology*. 3rd. Wellesley, MA: A K Peters Ltd.; 2006.
19. Abramowitz M, Stegun IA: *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications Inc.; 1966.
20. Wolfram Research I: *Mathematica, Version 9.01*. Champaign, Illinois: Wolfram Research, Inc.; 2013.

doi:10.1186/1471-2105-15-94
Cite this article as: Andrade *et al.*: The number of reduced alignments between two DNA sequences. *BMC Bioinformatics* 2014 **15**:94.