



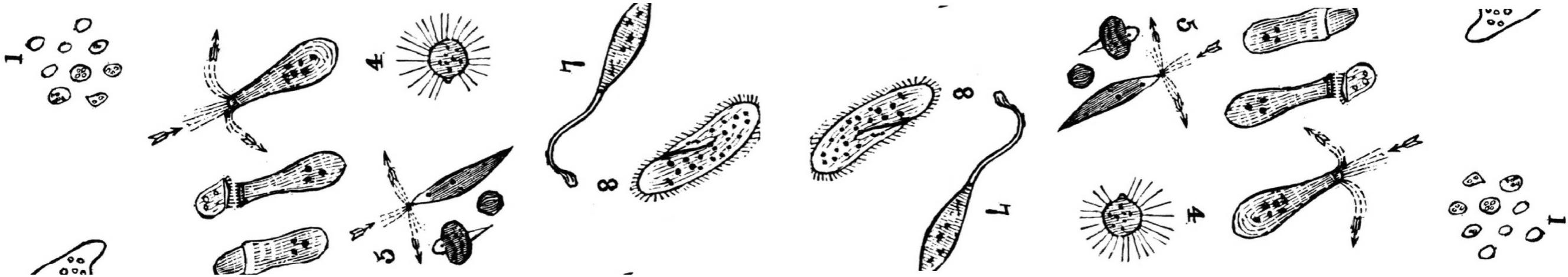
**CBIB**

CENTER FOR BIOINFORMATICS  
& INTEGRATIVE BIOLOGY

# Alineamiento de secuencias

[www.castrolab.org](http://www.castrolab.org)  
[www.cbib.cl](http://www.cbib.cl)

Eduardo Castro, PhD  
Universidad Andrés Bello  
9 de agosto de 2018



# ¿Qué es un alineamiento?

- Un alineamiento es una hipótesis de homología

# ¿Qué es una matriz de costo?

- Es una función matemática que determina el puntaje de un alineamiento.
- De entre muchos posibles alineamientos, los programas escogen el que tenga el mejor puntaje

# ¿Qué es una matriz de costo?

- Están definidas por una penalización sobre los “gaps”, extender “gaps”, y “mismatches”

ATTGACCTGA  
| | | | |  
AT - - CCTGA

Match = 1 punto  
Gap = -1 punto  
Total = 6

Match = 1 punto  
Abrir gap = -1 punto  
Extensión de gap = -1.5  
Total = 3

# ¿Qué es una matriz de costo?

- Están definidas por una penalización sobre los “gaps”, extender “gaps”, y “mismatches”

ACCTGATCCG	ACCTGATCCG
AC-TGATCAG	ACTGA-TCAG
S=8-4-3=1	S=5-4-12=-11

Figure 1.2. Alternate alignments of a pair of sequences illustrating a simple scoring function with matches = +1, mismatches = -3, and gaps = -4. The alignment on the left is better than the alignment on the right because its overall score is larger (1 vs. -11).

# ¿Es el alineamiento con mejor puntaje el verdadero?

- Usar genes bien caracterizados
- Simular datos (conoce la respuesta correcta)
- BaliBase → <http://www.lbgi.fr/balibase/>

The screenshot shows a journal article page. At the top left is a sidebar with icons for PDF (red square) and Info (green square with an 'i'). The main title is "Proteins: Structure, Function, and Bioinformatics" in orange, with a link "Explore this journal >". Below it is a "Research Article" section. The main title of the article is "BALiBASE 3.0: Latest developments of the multiple sequence alignment benchmark" in bold black text. Below the title are the authors' names: Julie D. Thompson, Patrice Koehl, Raymond Ripp, and Olivier Poch. Publication details at the bottom include "First published: 25 July 2005", "Full publication history", "DOI: 10.1002/prot.20527", and "View/save citation". To the right of the article text is a thumbnail of the journal cover for Volume 61, Issue 1, October 2005, with the word "PROTEINS" at the top. Below the cover are links: "View issue TOC", "Volume 61, Issue 1", "1 October 2005", and "Pages 127-136".

# ¿Qué puede pasar si tus genes están mal alineados?

- Búsqueda de motivos conservados
- Modelos de proteínas
- Árboles filogenéticos
- Selección positiva

Alineamiento local  
versus global

# Diferencias

## Alineamiento Local

alinear regiones locales de secuencias

secuencias relacionadas distamente; rearreglos; dominios compartidos

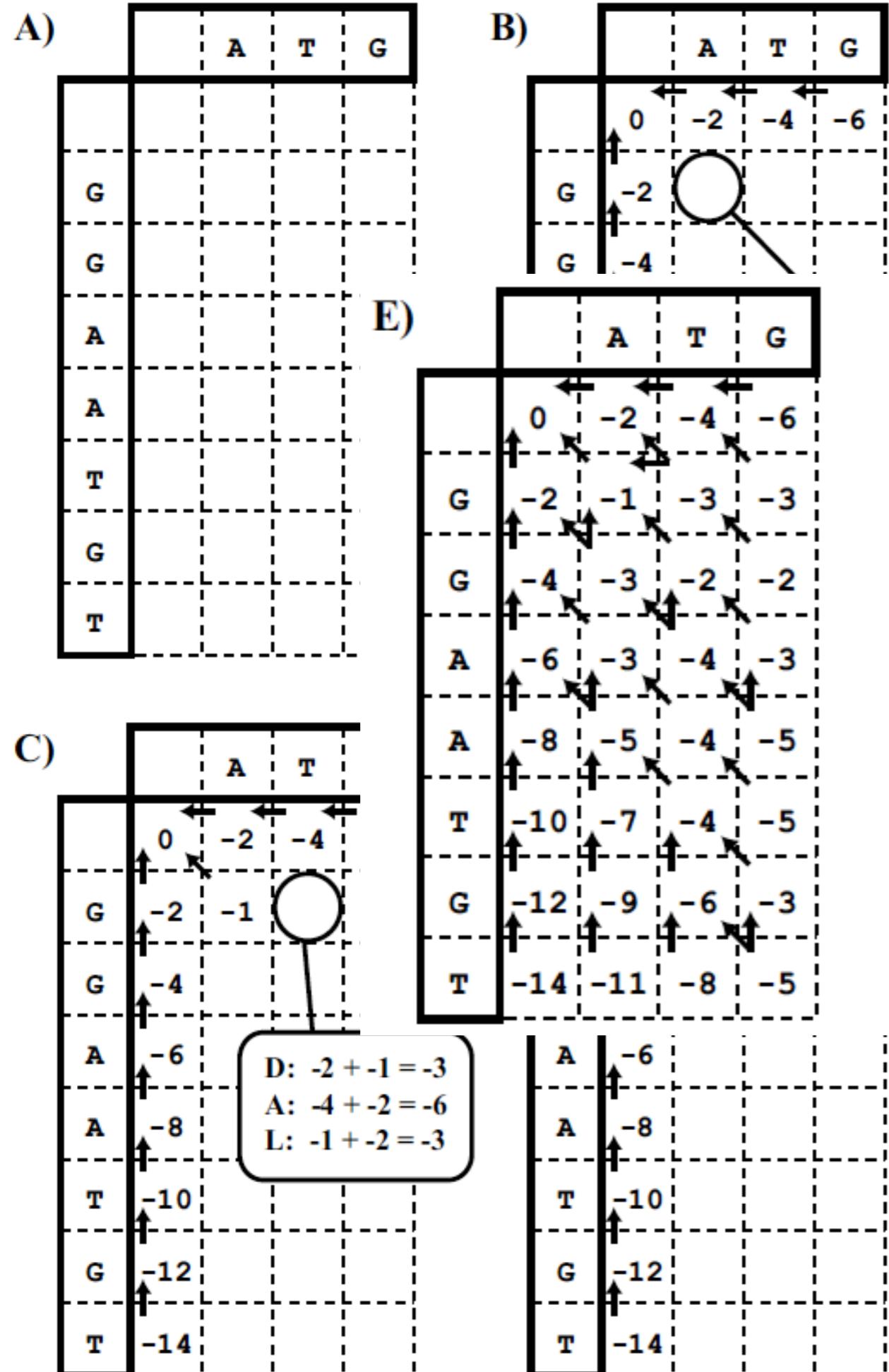
Algoritmo popular = Smith–Waterman

## Alineamiento Global

alinear dos secuencias de extremo a extremo

ideal para secuencias “cercanas” evolutivamente

Algoritmo popular = Needleman–Wunsch



# o global

match = +1  
mismatch = -1  
gap = -2

Figure 1.4. Illustration of Needleman-Wunsch (1970) global alignment algorithm. (A) Setting up the matrix. (B) The first row and column are filled with increasing multiples of the gap cost. The first cell will be given the maximum of three possible values. (C) The value for the first cell is entered along with the path that led to the value. The possible values for the second cell are illustrated. (D) The value for the second cell is entered; multiple paths are recorded since multiple paths led to the maximum score. (E) The completed matrix. (F) The completed matrix with all suboptimal paths removed. Tracing the arrows from the bottom right corner to the upper left leads to four possible paths and (therefore) four equally optimal alignments.

# Alineamiento global

GGAATGG  
---ATG-

GGAATGG  
---AT-G

GGAATGG  
--A-TG-

GGAATGG  
--A-T-G

Figure 1.5. Four equally optimal global alignments of sequences GGAATGG and ATG derived from the alignment matrix shown in Figure 1.2.

# Alineamiento global

- Existen secuencias que no pueden ser alineadas de extremo a extremo

AB--CDEF

ABEDC--F

ABCDEF

ABEDCF

ABCDE--F

AB--EDCF

Figure 1.6. Illustration of global alignment problem. Sequences ABCDEF and ABEDCF cannot be properly aligned because the homologous sections of the sequences are not in the same order.

# Alineamiento local

- Adaptación de Needleman-Wunsch —> permite un cuarto valor = 0

	A	T	G	
G	0	0	0	0
G	0	0	0	1
G	0	0	0	1
A	0	1	0	0
A	0	1	0	0
T	0	0	2	0
G	0	0	0	3
T	0	0	0	1

Figure 1.7. Completed score and trace-back matrix for local alignment using the Smith and Waterman (1981b) algorithm.

ATG  
ATG

# Comparación local/global

CAGCCTCGCTTAG  
AATGCCATTGACGG

A)

CA-GCC-TCGCTTAG	CA-GCC-TCGCTTAG	CA-GCC-TCGCTTAG	CA-GCC-TCGCTTAG
AATGCCATTGACG-G	AATGCCATTGAC-GG	AATGCCATTGA-CGG	AATGCCATTG-ACGG

B)

GCC  
GCC

+1 match, -1 mismatch, -2 gap

# ¿Cómo determinar qué valores usar?

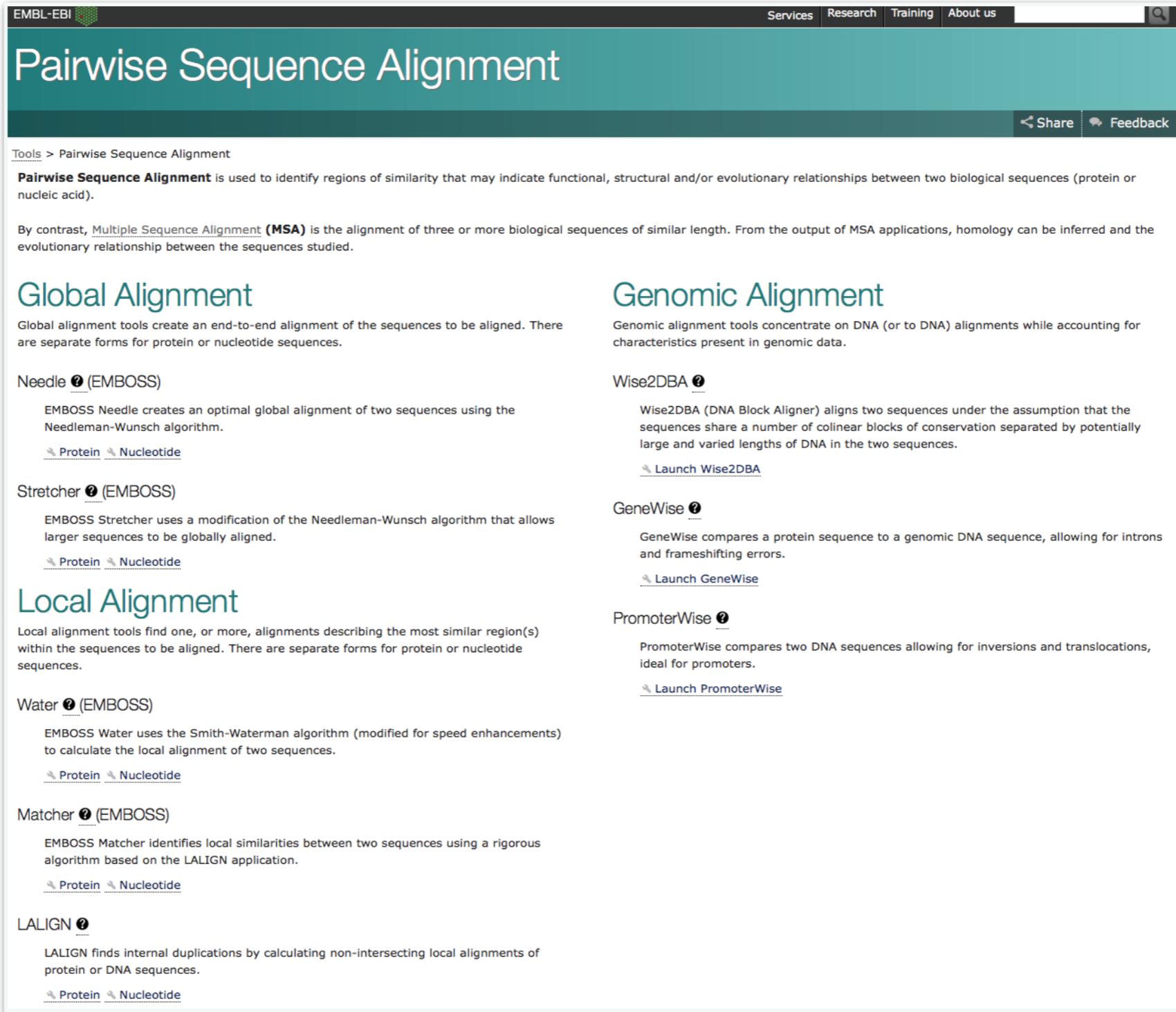
- La mayoría de la gente usa los valores “por defecto”, i.e., los que los autores pusieron
- proporción “mismatch”/“gap cost” lo más importante
- sustituciones/indels

GCC - TCG

GCCATTG

Figure 1.10. The optimal local alignment of sequences CAGCCTCGCTTAG and AATGCCATTGACGG with a cost function with matches = +1, mismatches = -0.3, and gaps = -1.3. Contrast with the local alignment in Figure 1.9B.

# Buen lugar para empezar = <http://www.ebi.ac.uk/Tools/psa/>



The screenshot shows the homepage of the EMBL-EBI Pairwise Sequence Alignment tool. The header includes the EMBL-EBI logo, navigation links for Services, Research, Training, and About us, and a search bar. Below the header, the page title is "Pairwise Sequence Alignment". A sub-navigation bar shows "Tools > Pairwise Sequence Alignment". The main content area is divided into two main sections: "Global Alignment" on the left and "Genomic Alignment" on the right.

**Global Alignment**

- Needle ? (EMBOSS)**: Describes the Needleman-Wunsch algorithm for global alignment of protein or nucleotide sequences. Includes links to "Protein" and "Nucleotide" versions.
- Stretcher ? (EMBOSS)**: Describes a modification of the Needleman-Wunsch algorithm for aligning larger sequences. Includes links to "Protein" and "Nucleotide" versions.
- Local Alignment**

  - Water ? (EMBOSS)**: Describes the Smith-Waterman algorithm for local alignment. Includes links to "Protein" and "Nucleotide" versions.
  - Matcher ? (EMBOSS)**: Describes a rigorous algorithm for identifying local similarities between sequences. Includes links to "Protein" and "Nucleotide" versions.
  - LALIGN ?**: Describes a tool for finding internal duplications by calculating non-intersecting local alignments. Includes links to "Protein" and "Nucleotide" versions.

**Genomic Alignment**

- Wise2DBA ?**: Describes Wise2DBA (DNA Block Aligner) for aligning DNA sequences. Includes a "Launch Wise2DBA" link.
- GeneWise ?**: Describes GeneWise for comparing protein sequences to genomic DNA. Includes a "Launch GeneWise" link.
- PromoterWise ?**: Describes PromoterWise for comparing DNA sequences, ideal for promoters. Includes a "Launch PromoterWise" link.

# Alineamiento múltiple

- Diferentes estrategias, e.g., matrices multidimensionales
- Alineamiento progresivo → Feng and Doolittle 1987

Article  
Journal of Molecular Evolution  
August 1987, Volume 25, Issue 4, pp 351-360

First online:

**Progressive sequence alignment as a prerequisite to correct phylogenetic trees**

Da-Fei Feng, Russell F. Doolittle



Look Inside >

# Alineamiento múltiple: pasos

1. Calcular todos los alineamientos de pares
  1. para  $n$  secuencias,  $n \times (n-1)/2$  pares
2. Calcular dendrograma usando algoritmo de clustering (UPGMA; Neighbor Joining)
3. Las secuencias más similares son alineadas primero de acuerdo al dendrograma

# Alineamiento múltiple: pasos

## - versión gráfica

calcular alineamientos de a pares para todas las secuencias

$S_1$  : PPGVKSDCAS  
 $S_2$  : PADGVVKDCAS  
 $S_3$  : PPDGKSDS  
 $S_4$  : GADGKDCCS  
 $S_5$  : GADGKDCAS



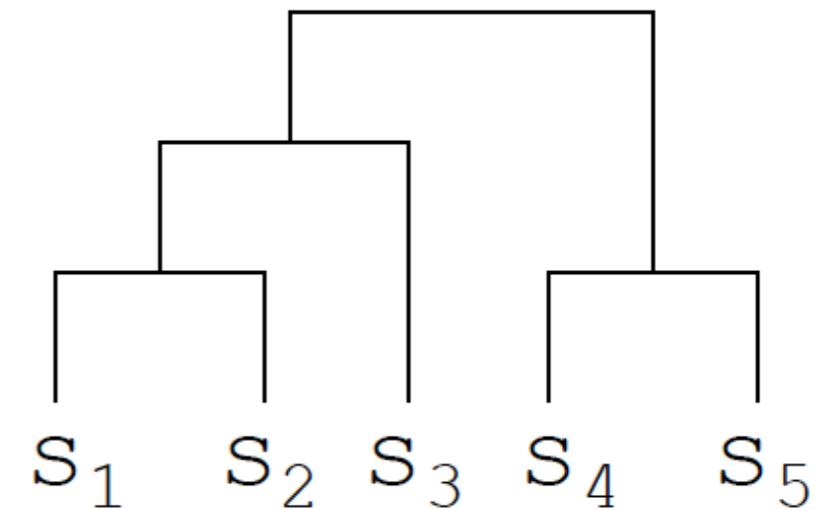
	$S_1$	$S_2$	$S_3$	$S_4$	$S_5$
$S_1$	0	0.111	0.25	0.555	0.444
$S_2$		0	0.375	0.222	0.111
$S_3$			0	0.5	0.5
$S_4$				0	0.111
$S_5$					0

# Alineamiento múltiple: pasos

## - versión gráfica

crear un dendrograma guía (árbol guía)

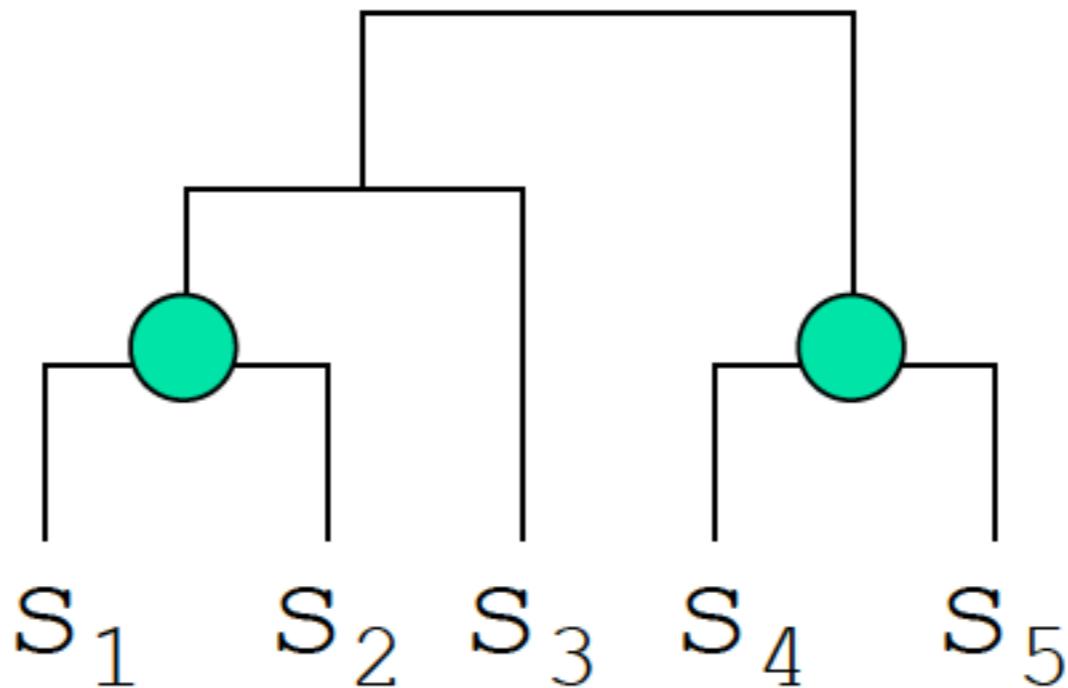
	S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>	S <sub>4</sub>	S <sub>5</sub>
S <sub>1</sub>	0	0.111	0.25	0.555	0.444
S <sub>2</sub>		0	0.375	0.222	0.111
S <sub>3</sub>			0	0.5	0.5
S <sub>4</sub>				0	0.111
S <sub>5</sub>					0



# Alineamiento múltiple: pasos

- versión gráfica

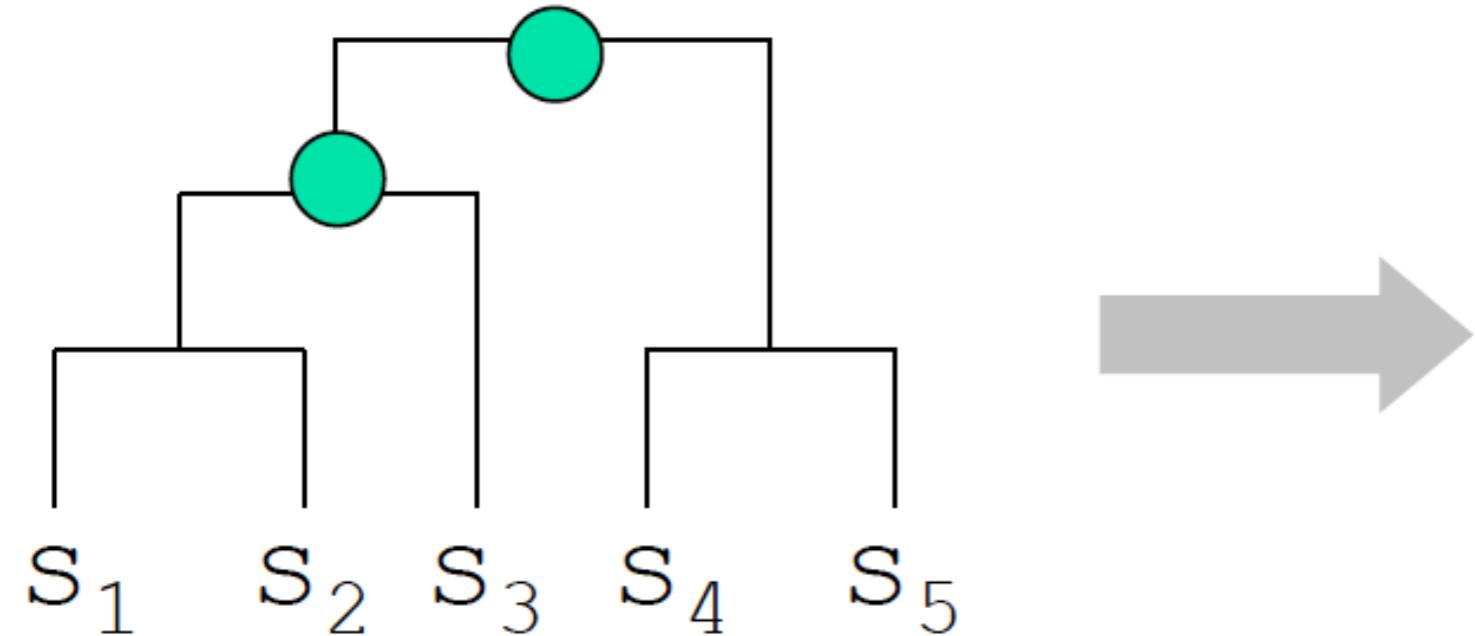
alineamos secuencias más similares primero



# Alineamiento múltiple: pasos

## - versión gráfica

Alineamiento múltiple



$S_1$ : P-PGVKSDCAS  
 $S_2$ : PADGVK-DCAS  
 $S_3$ : PPDG-KSD--S  
 $S_4$ : GADG-K-DCCS  
 $S_5$ : GADG-K-DCAS

# Alineamiento múltiple: otros métodos

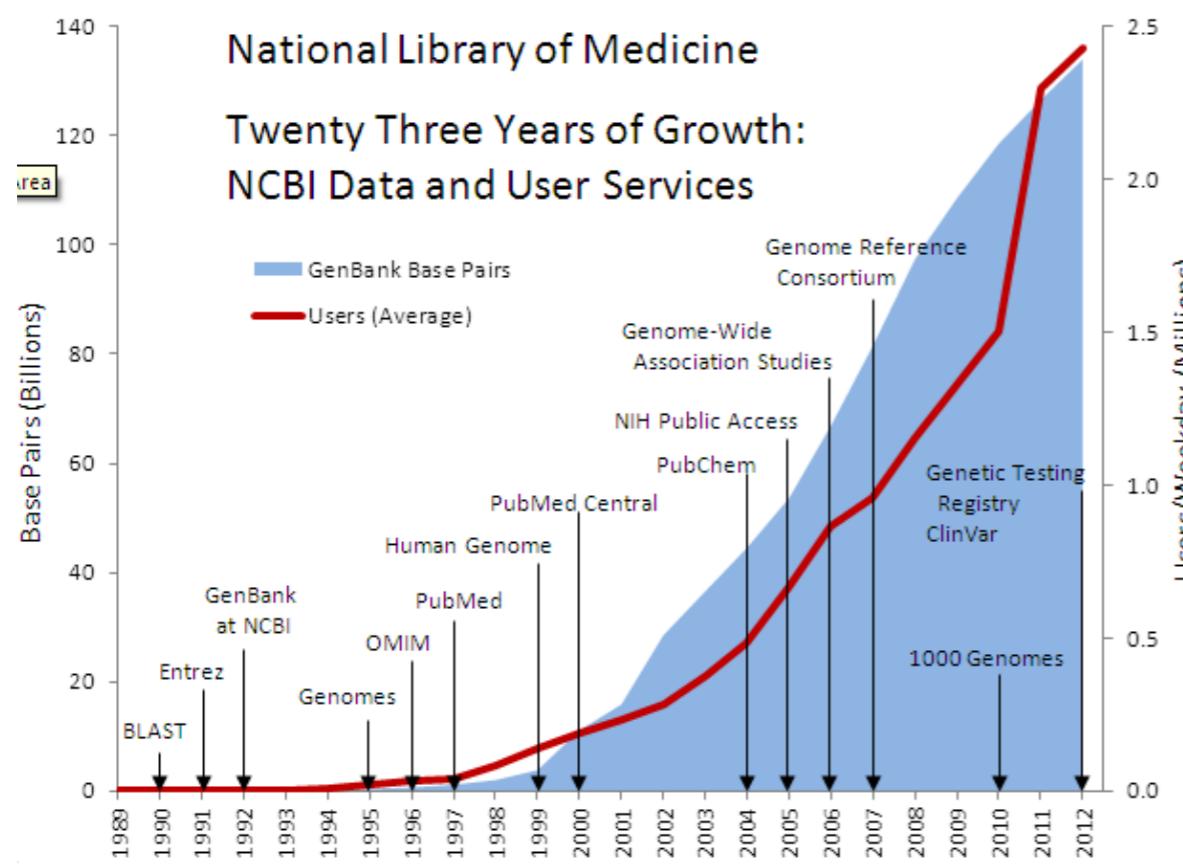
- métodos iterativos —> Muscle [http://www.ebi.ac.uk/  
Tools/msa/muscle/](http://www.ebi.ac.uk/Tools/msa/muscle/)
- métodos de consenso —> M-Coffee [http://  
www.tcoffee.org/Projects/mcoffee/](http://www.tcoffee.org/Projects/mcoffee/)
- modelos ocultos de Markov —> HMMER [http://  
hmmer.org](http://hmmer.org)
- algoritmos genéticos —> SAGA [http://  
www.tcoffee.org/Projects/saga/](http://www.tcoffee.org/Projects/saga/)

# BLAST

## Basic Local Alignment Search Tool

# BLAST

- Tipo de alineamiento local especializado para búsqueda en bases de datos
- En segundos entrega resultados



Genetic Sequence Data Bank  
February 15 2016  
190250235 loci, 207018196067 bases,  
from 190250235 reported sequences

# BLAST

- Entrega resultados que no se deben a chance
- La premisa es que si dos secuencias se parecen no por chance, entonces son homólogas
- Homología: Desciende de un ancestro común, no necesariamente la misma función

Program	Description
<b>blastp</b>	Compares an amino acid query sequence against a protein sequence database.
<b>blastn</b>	Compares a nucleotide query sequence against a nucleotide sequence database.
<b>blastx</b>	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
<b>tblastn</b>	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
<b>tblastx</b>	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Program	Notes
Megablast	<p><b>Contiguous</b></p> <p><b>Discontiguous</b></p>
Position Specific	<p><b>PSI-BLAST</b></p> <p><b>RPS-BLAST</b></p>



nucleotide only



protein only

# ¿Cómo funciona?

- Matriz de costo
- Corta tu secuencia en fragmentos de 3 nucleótidos
- Esos fragmentos (semillas) son extendidos hasta encontrar un resultado óptimo

# Las estadísticas de BLAST son importantes

- Te permiten distinguir entre resultados significativos y por chance
- Los principales son el Score (puntaje), Query Coverage (cobertura de la secuencia de consulta) y el e-value

# Las estadísticas de BLAST son importantes\*

- Score = La suma de los puntajes para cada posición en la secuencia de consulta y su resultado → **Representa la calidad del alineamiento**
- e-value = No es una probabilidad, es una expectativa. **Representa el número de alineamientos diferentes con puntajes equivalentes o mejores que los que se esperan ocurrían por chance**

\*pregunta de prueba

# e-value

- **Representa el número de alineamientos diferentes con puntajes equivalentes o mejores que los que se esperan ocurrían por chance**
- e-value bajo ( $10^{-5}$ ) puede ser indicativo de homología
- Excepción: regiones de baja complejidad como repeticiones, pueden tener e-value bajo pero no ser homólogas

BLAST® Basic Local Alignment Search Tool My NCBI Welcome pevsner. [Sign Out]

NCBI/ BLAST/ blastp suite Standard Protein BLAST

blastn blastp blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#) Reset page Bookmark

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

>gi|4504349|ref|NP\_000509.1| hemoglobin subunit beta [Homo sapiens]  
MVLHTPEEKSAVTALWKGKVNVDEVGGEALGRLLVVYWPWTQRFESFGDLSPTPDAVMGNPKVKAH  
GKKVLGAFSDGLAHLDNLKGTFAFLSELHCDKLHVDPENFRLGNVLVCVLAAHFGKEFTPVQ  
AYQKVVAGVANALAHKYH

Clear Query subrange [?](#)

From \_\_\_\_\_ To \_\_\_\_\_

Or, upload file  [Browse...](#) [?](#)

Job Title  Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**Choose Search Set**

Database  Reference proteins (refseq\_protein) [?](#)

Organism Optional   Exclude [+](#)  
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude Optional  Models (XM/XP)  Uncultured/environmental sample sequences

Entrez Query Optional  Enter an Entrez query to limit search [?](#)

**Program Selection**

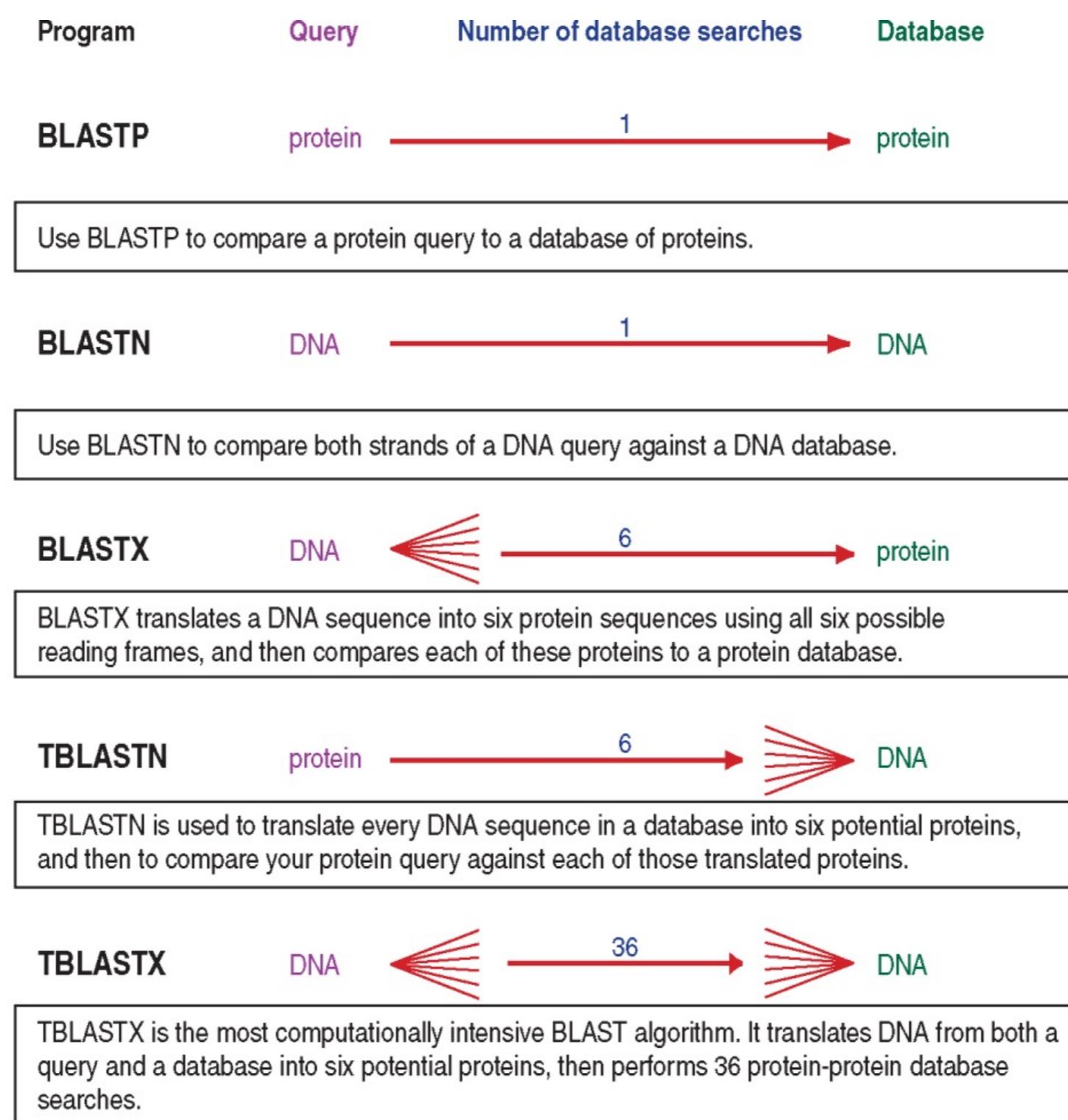
Algorithm  blastp (protein-protein BLAST)  
 PSI-BLAST (Position-Specific Iterated BLAST)  
 PHI-BLAST (Pattern Hit Initiated BLAST)  
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)  
 Choose a BLAST algorithm [?](#)

**BLAST** Search database Reference proteins (refseq\_protein) using Blastp (protein-protein BLAST)  
 Show results in a new window

**Algorithm parameters** Note: Parameter values that differ from the default are highlighted

## Sub-unidad beta de la hemoglobina

**FIGURE 4.1** Main page for a BLASTP search at NCBI. The sequence can be input as an accession number, GI identifier, or FASTA-formatted sequence as shown here (arrow 1). The database must be selected (arrow 2) if the default setting is not selected (as here, in which the database is set to RefSeq proteins); the choice is highlighted in yellow. The search can be restricted to a particular organism or taxonomic group, and Entrez queries can be used to further focus the search (arrow 3); here we limit the search to entries including the author Max Perutz. We discuss the BLASTP algorithm in this chapter (arrow 4), and PSI-BLAST, PHI-BLAST, and DELTA-BLAST in Chapter 5. Many of the search parameters can be modified (arrow 5).

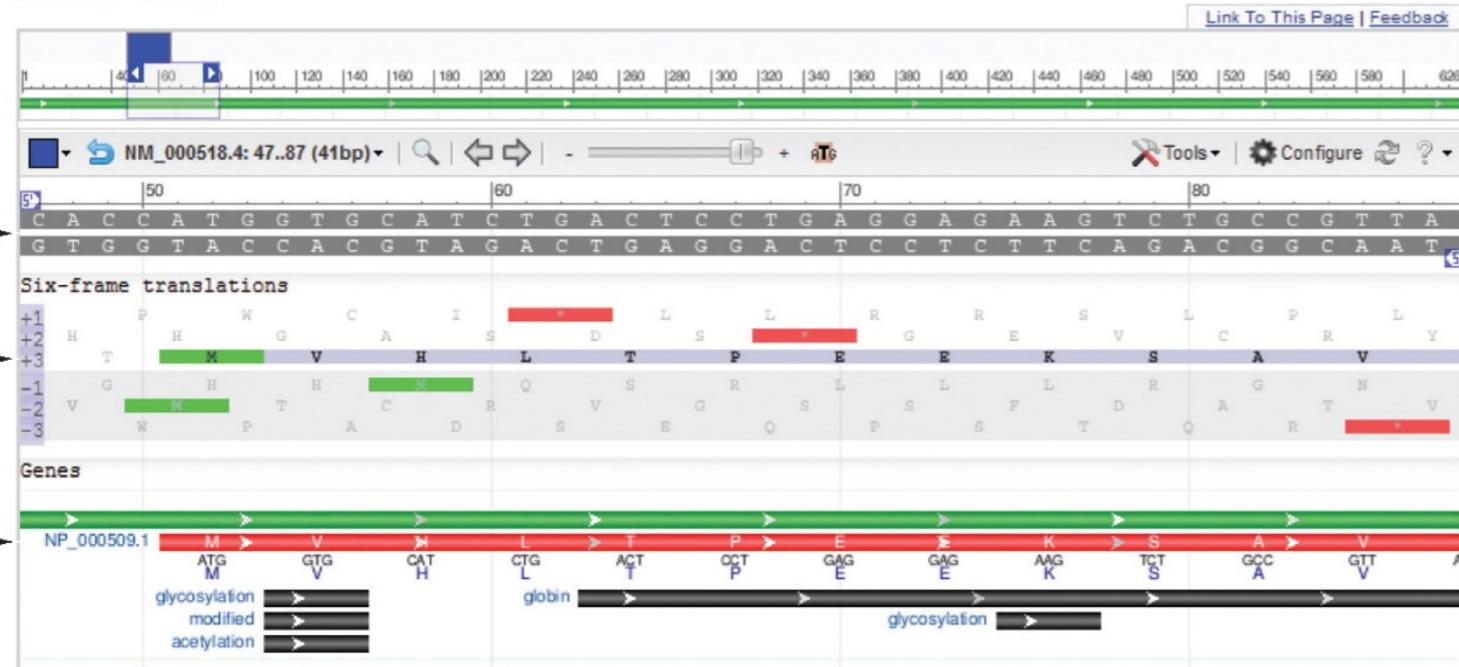


**FIGURE 4.2** Overview of the five main BLAST algorithms. Note that the suffix P refers to protein (as in BLASTP), N refers to nucleotide, and X refers to a DNA query that is dynamically translated into six protein sequences. The prefix T refers to “translating,” in which a DNA database is dynamically translated into six proteins.

## Homo sapiens hemoglobin, beta (HBB), mRNA

NCBI Reference Sequence: NM\_000518.4

[GenBank](#) [FASTA](#)



¿Cuántos marcos de lectura existen?

**FIGURE 4.3** DNA can potentially encode six different proteins. To demonstrate this, we view the NCBI Nucleotide entry for HBB and select the “graphics” view; The two strands of DNA sequence are shown (arrow 1). In this zoomed view, only a portion of the HBB sequence is displayed. From the top strand, three potential proteins are encoded (frames +1, +2, +3) with the corresponding amino acids indicated in gray using the single-letter amino acid abbreviations. In this case, frame +3 corresponds to the frame used for translation (arrow 2). Note that frames +1 and +2 as well as frame -3 include stop codons (asterisks shaded red). The lower portion of the display includes the amino acid sequence of the corresponding protein (arrow 3) as well as the corresponding nucleotides (matching frame +3); features indicated with black shading represent a site that may be acetylated or glycosylated and a globin domain.

**Algorithm parameters**

**General Parameters**

1 → Max target sequences 100 Select the maximum number of aligned sequences to display ?

2 → Short queries  Automatically adjust parameters for short input sequences ?

3 → Expect threshold 10 ?

4 → Word size 3 ?

5 → Max matches in a query range 0 ?

**Scoring Parameters**

6 → Matrix BLOSUM62 ?

7 → Gap Costs Existence: 11 Extension: 1 ?

8 → Compositional adjustments Conditional compositional score matrix adjustment ?

**Filters and Masking**

9 → Filter  Low complexity regions ?

10 → Mask  Mask for lookup table only ?  
 Mask lower case letters ?

**BLAST** Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)  
 Show results in a new window

**FIGURE 4.4** Optional BLASTP parameters. Numbered arrows refer to discussion in the text.

*Bioinformatics and Functional Genomics*, Third Edition, Jonathan Pevsner.

© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.

Companion Website: [www.wiley.com/go/pevsnerbioinformatics](http://www.wiley.com/go/pevsnerbioinformatics)

## Graphic Summary

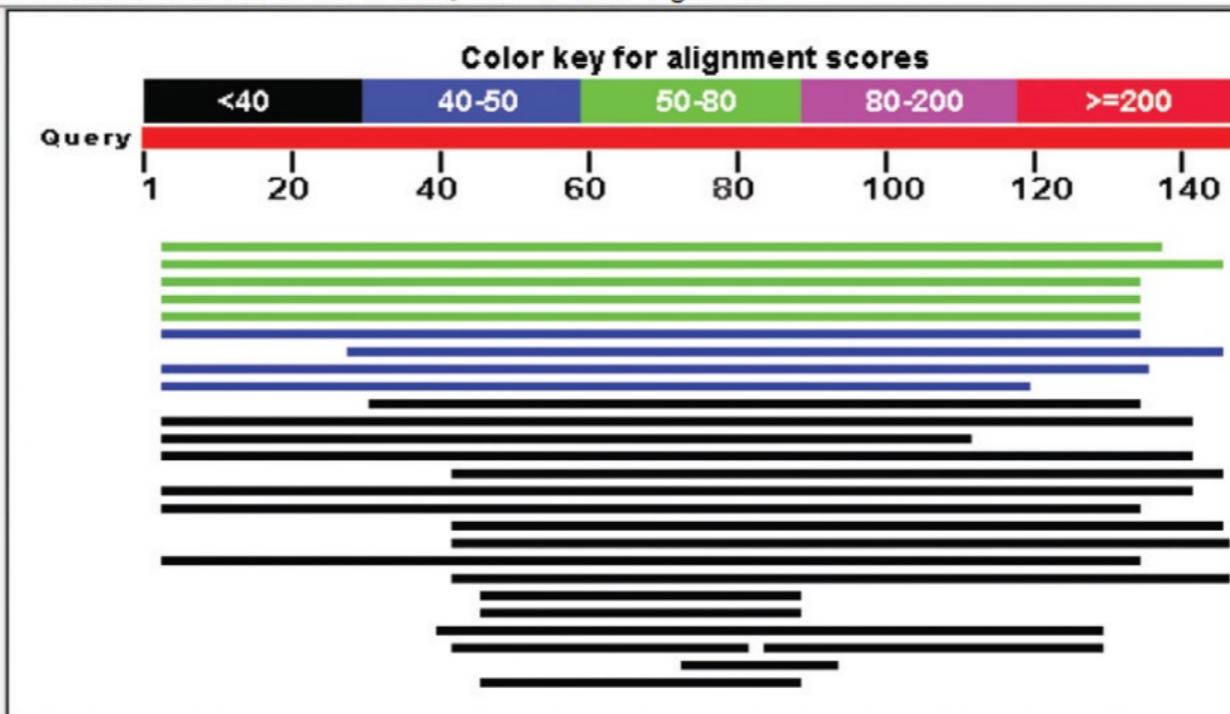
Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 27 Blast Hits on the Query Sequence ⓘ

Mouse-over to show define and scores, click to show alignments



**FIGURE 4.8** The graphic summary of BLAST results includes a display of conserved domains (here showing a match to the globin protein family), then a color-coded distribution of hits. Here the *x* axis corresponds to the length of the query (147 amino acid residues for beta globin), with each database match characterized by a color-coded score (e.g., five matches shaded green have scores of 50–80) and lengths (one of the five green database hits includes an aligned region that extends fully to the carboxy-terminus of the HBB query, while the other four do not). This graphic can be useful to summarize the regions in which database matches align to the query.

## Sequences producing significant alignments:

Select: All None Selected:2

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input checked="" type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396833.1  PREDI	59.7	59.7	91%	1e-10	29%	XP_003396832.1
<input checked="" type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus impatiens] >ref XP_003494220.1  PREDI	58.5	58.5	97%	3e-10	28%	XP_003494219.1
<input type="checkbox"/>	PREDICTED: globin-like [Megachile rotundata]	57.8	57.8	89%	6e-10	29%	XP_003707185.1
<input type="checkbox"/>	PREDICTED: globin-like [Apis florea]	53.9	53.9	89%	1e-08	30%	XP_003690810.1
<input type="checkbox"/>	globin 1 [Apis mellifera]	52.8	52.8	89%	4e-08	30%	NP_001071291.1
<input type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396831.1  PREDI	45.1	45.1	89%	2e-05	26%	XP_003396830.1
<input type="checkbox"/>	PREDICTED: neuroglobin-like, partial [Acyrthosiphon pisum]	42.4	42.4	80%	2e-04	23%	XP_001946608.2
<input type="checkbox"/>	globin, putative [Ixodes scapularis]	42.7	42.7	90%	2e-04	25%	XP_002414906.1

**FIGURE 4.9** A typical BLASTP output includes a list of database sequences that match the query. Links are provided to that database entry (e.g., an NCBI Protein entry) and to the pairwise alignment to the query. The bit score and *E* value for each alignment are also provided. Note that the best matches at the top of the list have large bit scores and small *E* values.

# Estrategias para mejorar alineamientos

- Si tienes secuencias codificantes → usa un alineamiento traducido (translated alignment)

Direct nucleotide alignment

A color-coded alignment of nine DNA sequences. Nucleotides are represented by four colors: orange for A, green for C, blue for G, and red for T. The alignment shows high conservation of sequence across the samples.

Average % of identity: 78%  
(ranging from 57 to 88%)

Amino acid alignment

A color-coded alignment of nine protein sequences. Amino acids are represented by four colors: orange for K, green for F, blue for M, and red for I. The alignment shows high conservation of sequence across the samples.

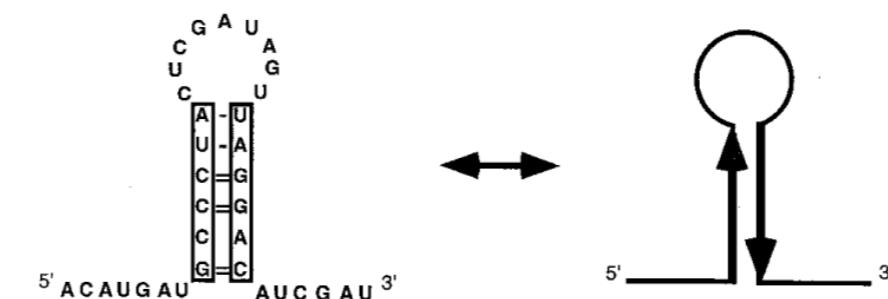
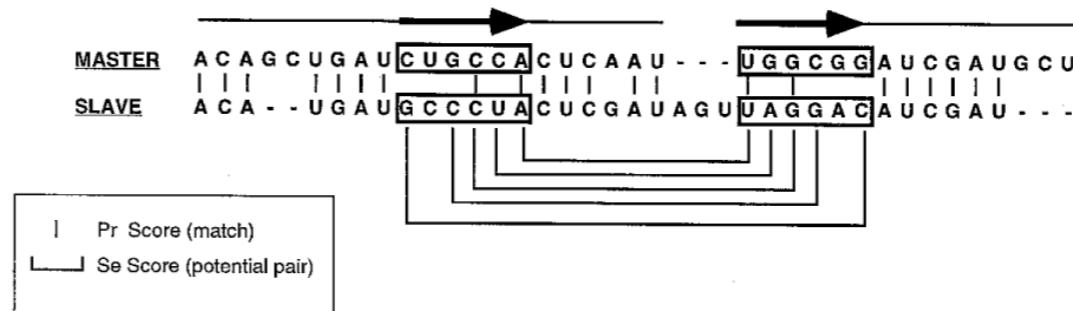
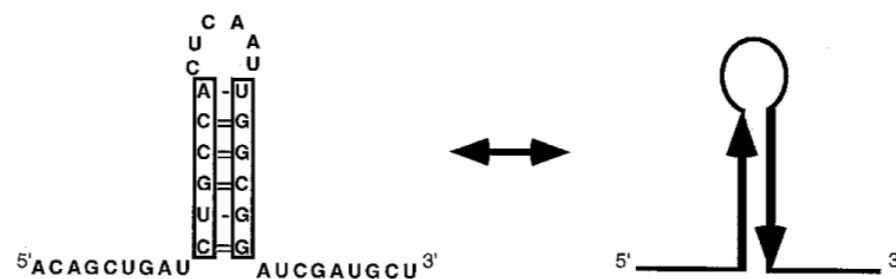
Back-translation

Average % of identity: 73%  
(ranging from 33 to 88%)

A color-coded alignment of nine DNA sequences, representing the back-translation of the amino acid alignment. The nucleotide colors (orange, green, blue, red) correspond to the amino acid colors (K, F, M, I). The alignment shows high conservation of sequence across the samples.

# Estrategias para mejorar alineamientos

- Si tienes secuencias que forman estructuras secundarias —> usa un alineamiento ad hoc (structure-aware)



Ahora el laboratorio