Alineamiento de secuencias

Genómica para bioinformática INB320
29 marzo 2016
Eduardo Castro-Nallar, PhD
www.castrolab.org

Todo en bioinformática es comparativo

- …y prácticamente todo comienza con un alineamiento de secuencias

- Alinear secuencias es comúnmente un procedimiento al cual los investigadores no le prestan mucha atención…

Todo en bioinformática es comparativo

Aplicación	Objetivo
Extrapolación	Asignar una secuencia no caracterizada a una familia de proteínas
Análisis filogenético	Reconstrucción de la historia de una proteína o familia de proteínas
Identificación de patrones	Identificación de regiones características de una función por identificación de sitios conservados
Identificación de dominios	Generación de perfiles específicos para identificar nuevos dominios en secuencias no caracterizadas
Elementos regulatorios de ADN	Generar una matriz de peso a partir de un alineamiento para escanear otras secuencias de ADN e identificar sitios de unión
Predicción de estructura	Predecir estructura secundaria y ayudar a generar modelos 3D
PCR	Identificación de regiones conservadas para generar partidores para PCR

www.cs.uml.edu/~grinstei/91.510/msa_ppt.ppt

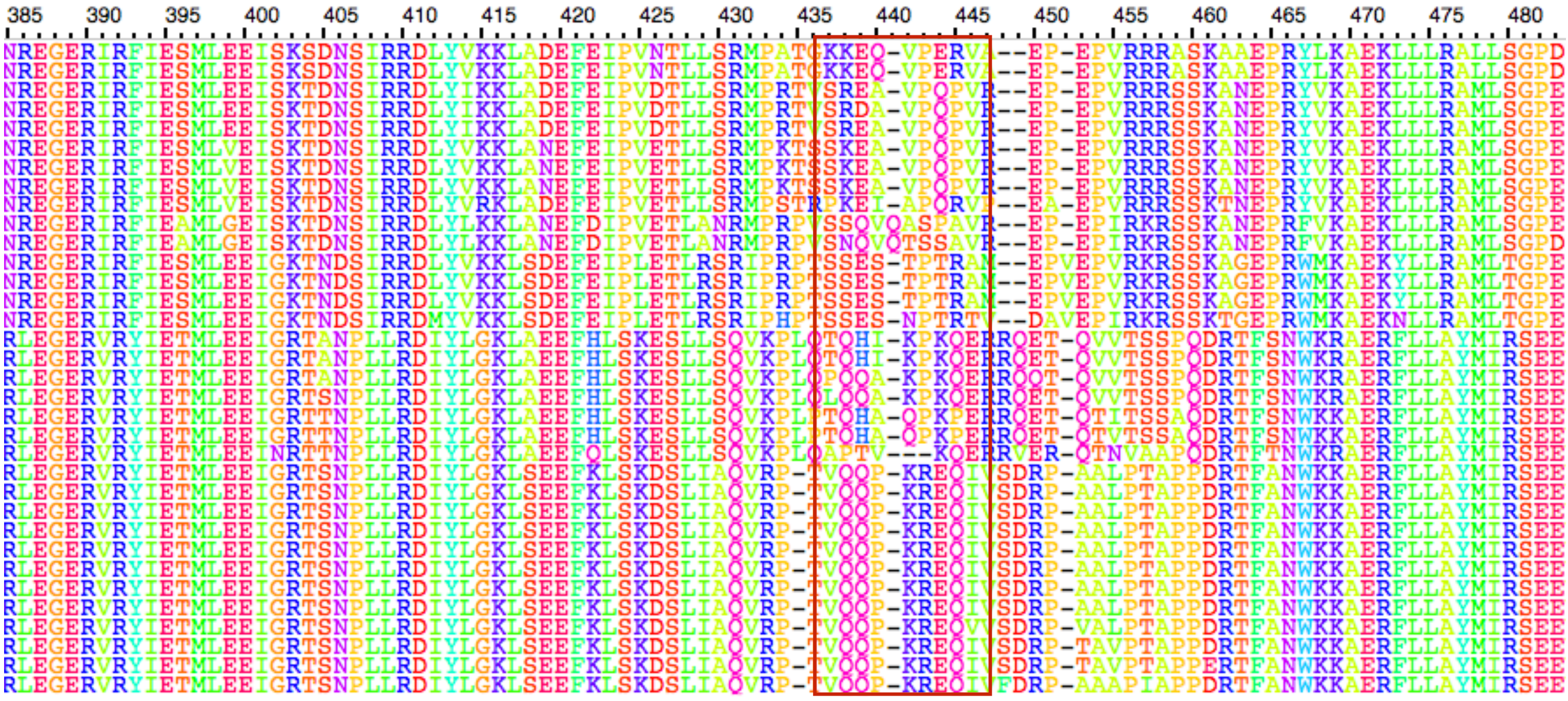
Todo en bioinformática es comparativo

- … sin embargo las consecuencias de los errores de alineamiento son arrastrados durante todo un análisis

- equivalente en biología molecular —> equilibrar mal el pH de una solución

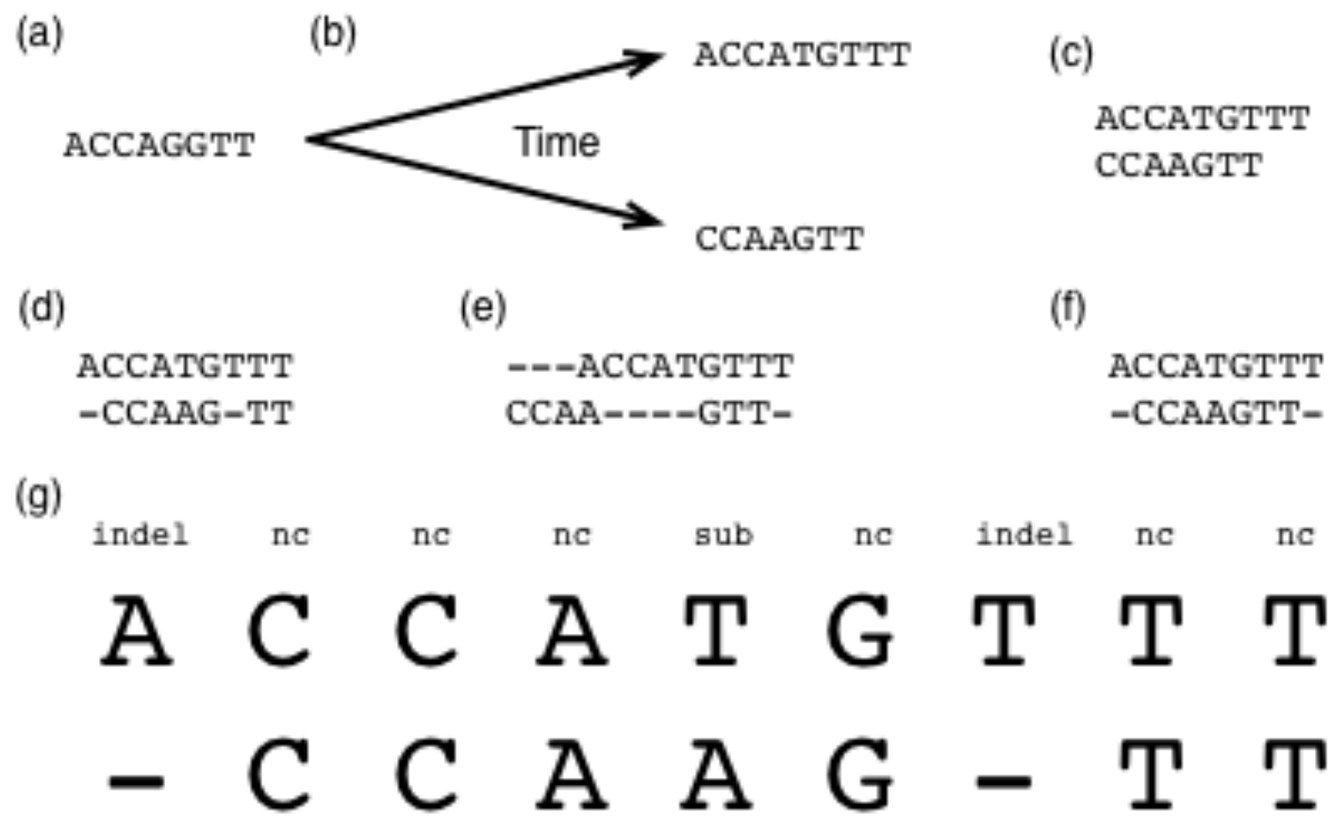
¿Qué es un alineamiento?

- Un alineamiento es una hipótesis de homología



¿Qué es un alineamiento?

- Un alineamiento es una hipótesis de homología



The goal of pairwise sequence alignment is, given two sequences, to generate a hypothesis about which sequence positions derived from a common ancestral sequence position

http://readiab.org/book/latest/

¿Qué es un alineamiento?

- Principio de parsimonia - navaja de Occam
- “Among competing hypotheses, the one with the fewest assumptions should be selected.” Wikipedia
- En ciencia, es una heurística para desarrollar modelos —> en alineamientos = maximizar la similitud entre dos secuencias introduciendo el mínimo de mutaciones o las menos extremas

http://jbwarner86.deviantart.com/art/F16-Occam-s-Razor-183519777

Errores en alineamientos dependen de grado de divergencia

- Confiar ciegamente en un programa —> biología molecular vs. biología computacional

"Garbage in - garbage out"

Identifying errors in sequence alignment to improve protein comparative modelling

Danielle Talbot† and Andrew C.R. Martin*

RESEARCH ARTICLE Open Access

Characterization of multiple sequence alignment errors using complete-likelihood score and position-shift map

Kiyoshi Ezawa[1,2]

Contents lists available at ScienceDirect

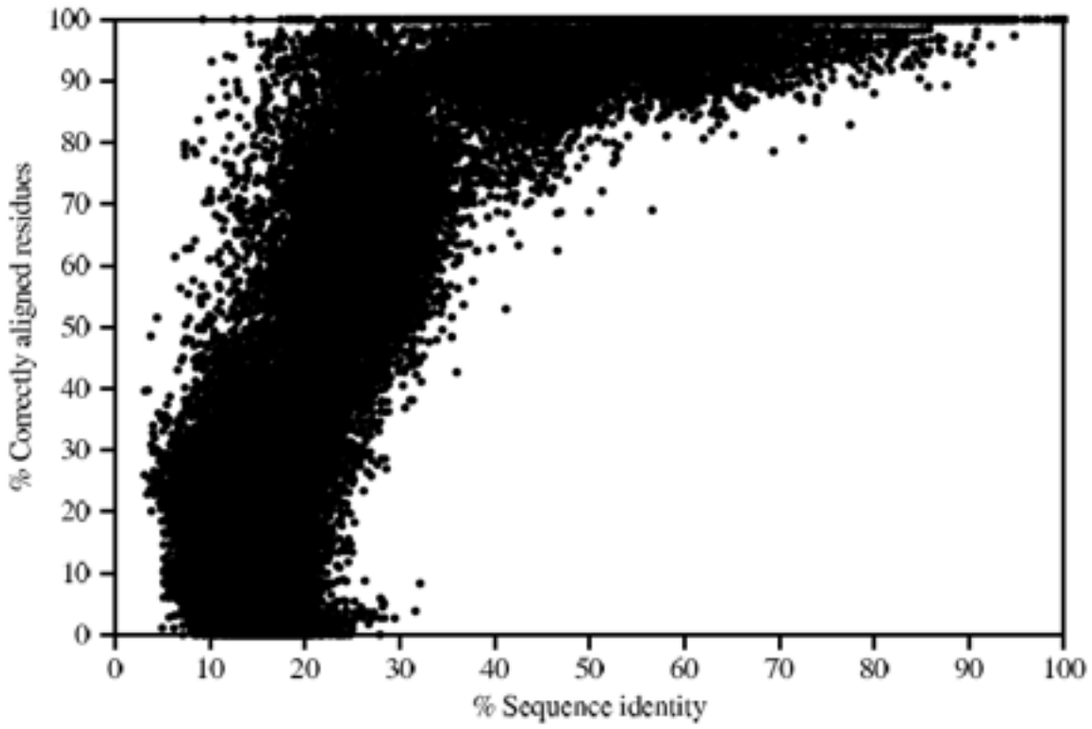
Gene

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

Characterization of pairwise and multiple sequence alignment errors

Giddy Landan *, Dan Graur

Department of Biology & Biochemistry, University of Houston, Houston, TX, USA



Errores en alineamientos

Contents lists available at ScienceDirect

Gene

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

Characterization of pairwise and multiple sequence alignment errors

Giddy Landan *, Dan Graur

Department of Biology & Biochemistry, University of Houston, Houston, TX, USA

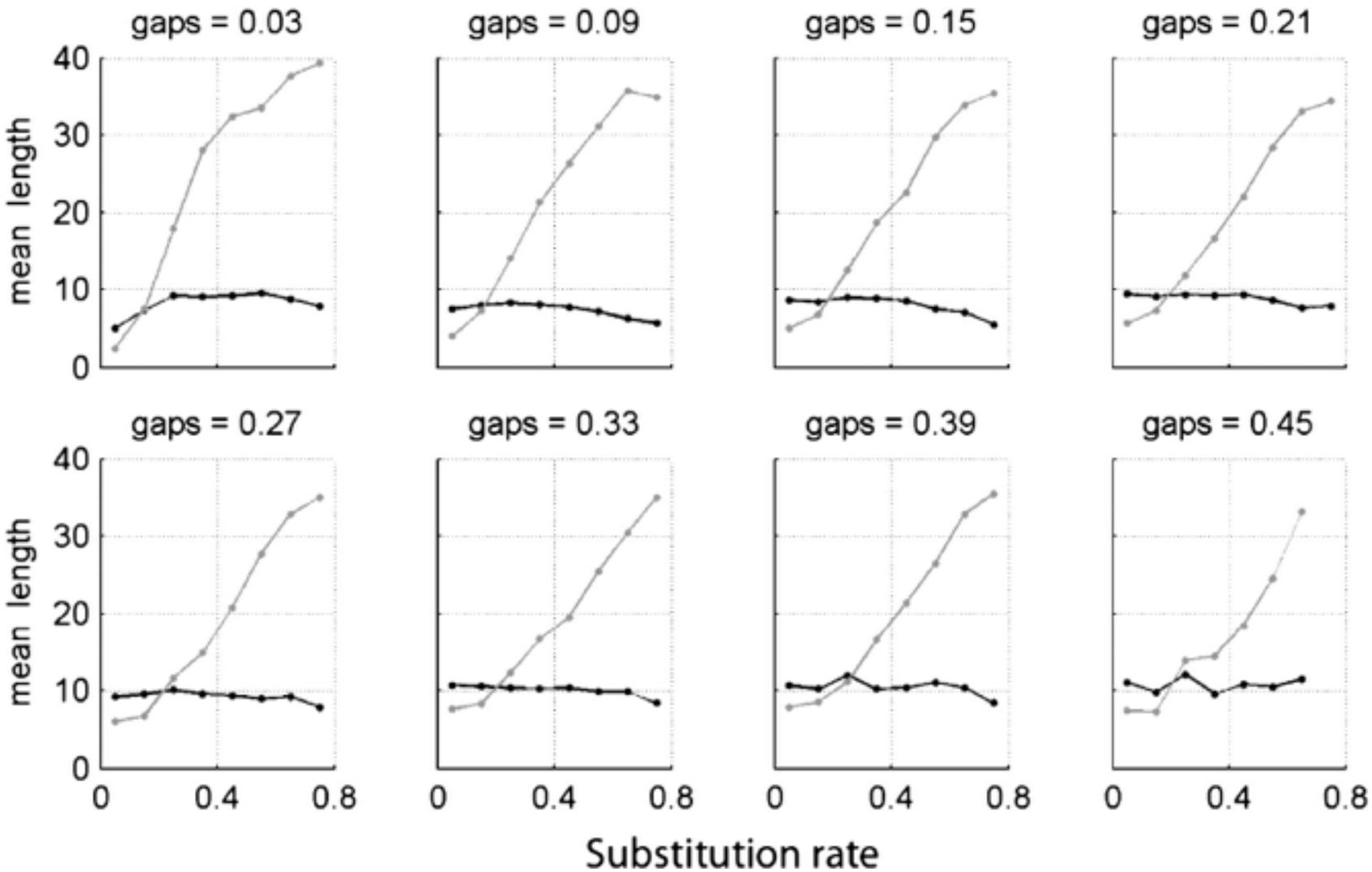
- "Error rates rapidly increase with **sequence divergence**, thus, for even intermediate degrees of sequence divergence, **more than half of the columns** of a reconstructed alignment may be **expected to be erroneous**"

- "**In closely related sequences**, most errors consist of the erroneous positioning of a single indel event and **their effect is local**"

- "**Correct reconstruction** can only be guaranteed when the likelihood of true alignment is uniquely **optimal**. However, true alignment features are very frequently **sub-optimal or co-optimal**, with the result that optimal albeit **erroneous features are incorporated** into the reconstructed MSA"

Landan, G. and Graur, D., 2009. Characterization of pairwise and multiple sequence alignment errors. *Gene*, *441*(1), pp.141-147.

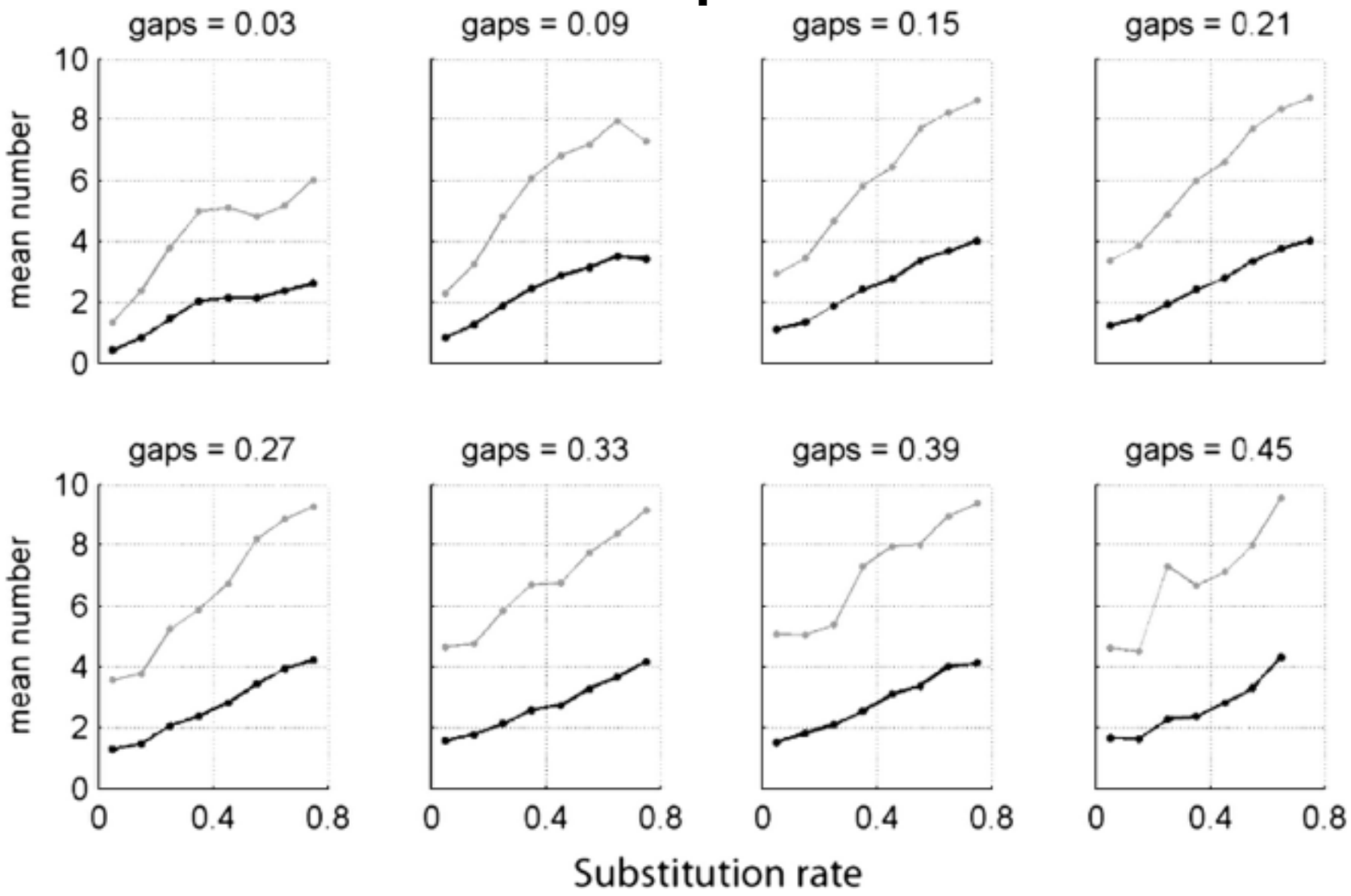
Errores en alineamientos pairwise o de pares - sustituciones



Next, we note that the mean length of error segments (Fig. 3, gray lines) increases dramatically with substitution rate, while the mean length of correctly reconstructed segments remains fairly stable (Fig. 3, black lines). While the length of error segments increase with divergence, we note that **erroneously reconstructed segments contain fewer indels** (and gap characters) and are shorter than the corresponding true segments. This is a systematic bias resulting from the strict optimization of the objective function coupled with the fact that, for the same number of matches, shorter alignments usually score better than longer ones.

Landan, G. and Graur, D., 2009. Characterization of pairwise and multiple sequence alignment errors. *Gene*, *441*(1), pp.141-147.

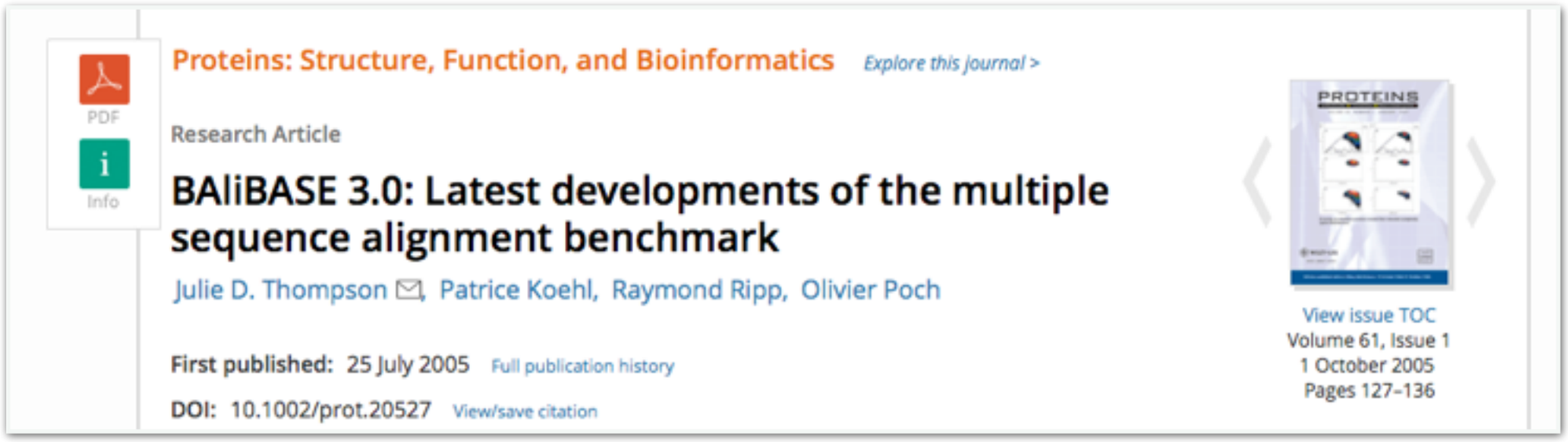
Errores en alineamientos pairwise o de pares - indels



The **mean numbers of wrongly inferred indels and gap-character states increases with substitution rate** (Fig. 4). For **closely related sequences**, the error segments are short and frequently result from a **single indel being erroneously positioned**. As the two sequences farther **diverge**, the **errors multiply.**

Evaluación de métodos

- Usar genes bien caracterizados
- Simular datos (conoce la respuesta correcta)
- BaliBase —> http://www.lbgi.fr/balibase/
- Otros “benchmarks” = PREFAB, SABMARK, OXBENCH y IRMBASE



Ahora que sabemos qué es un alineamiento y que todos los alineamientos están malos, revisemos cómo se construyen

¿Qué es una matriz de costo?

Cost matrix/objective function/Scoring matrix

- 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

- No existe una matriz que funcione para todas las secuencias que una persona quisiera alinear. “Adivinar” basándose en % de identidad

¿Qué es una matriz de costo?

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

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).

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

Alineamiento local versus global

Diferencias

Alineamiento Local	Alineamiento Global
alinear regiones locales de secuencias	alinear dos secuencias de extremo a extremo
secuencias relacionadas distantemente; rearreglos; dominios compartidos	ideal para secuencias “cercanas” evolutivamente
Algoritmo popular = Smith–Waterman	Algoritmo popular = Needleman–Wunsch

Alineamiento 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.

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

o global

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

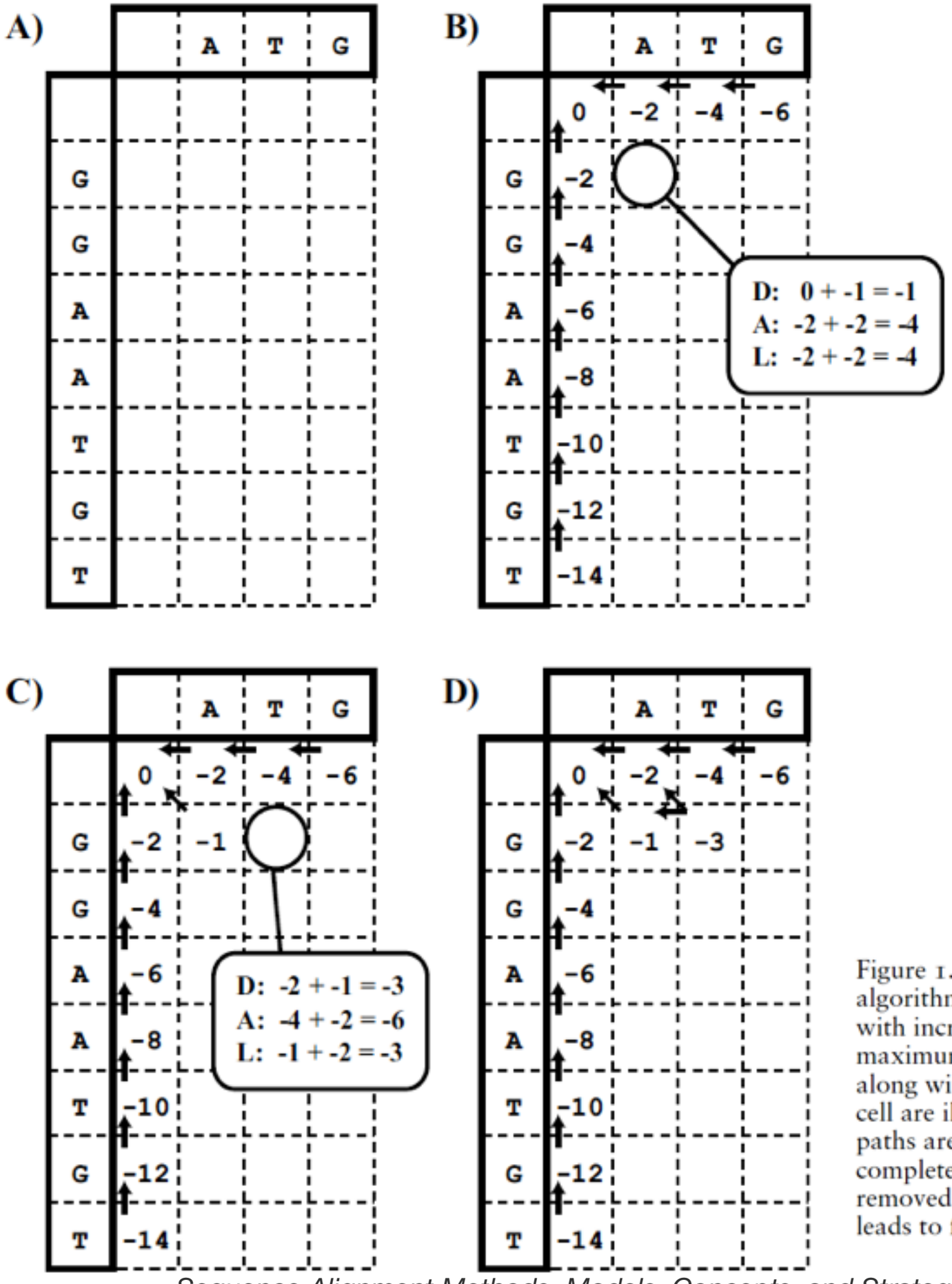


Figure 1.4. Illustration of Needleman–Wunsch (1970) global alignment algorithm. (A) Setting up the matrix. (B) The first row and column are filled with increasing multiples of the gap cost. The first cell will be given the maximum of three possible values. (C) The value for the first cell is entered 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.

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

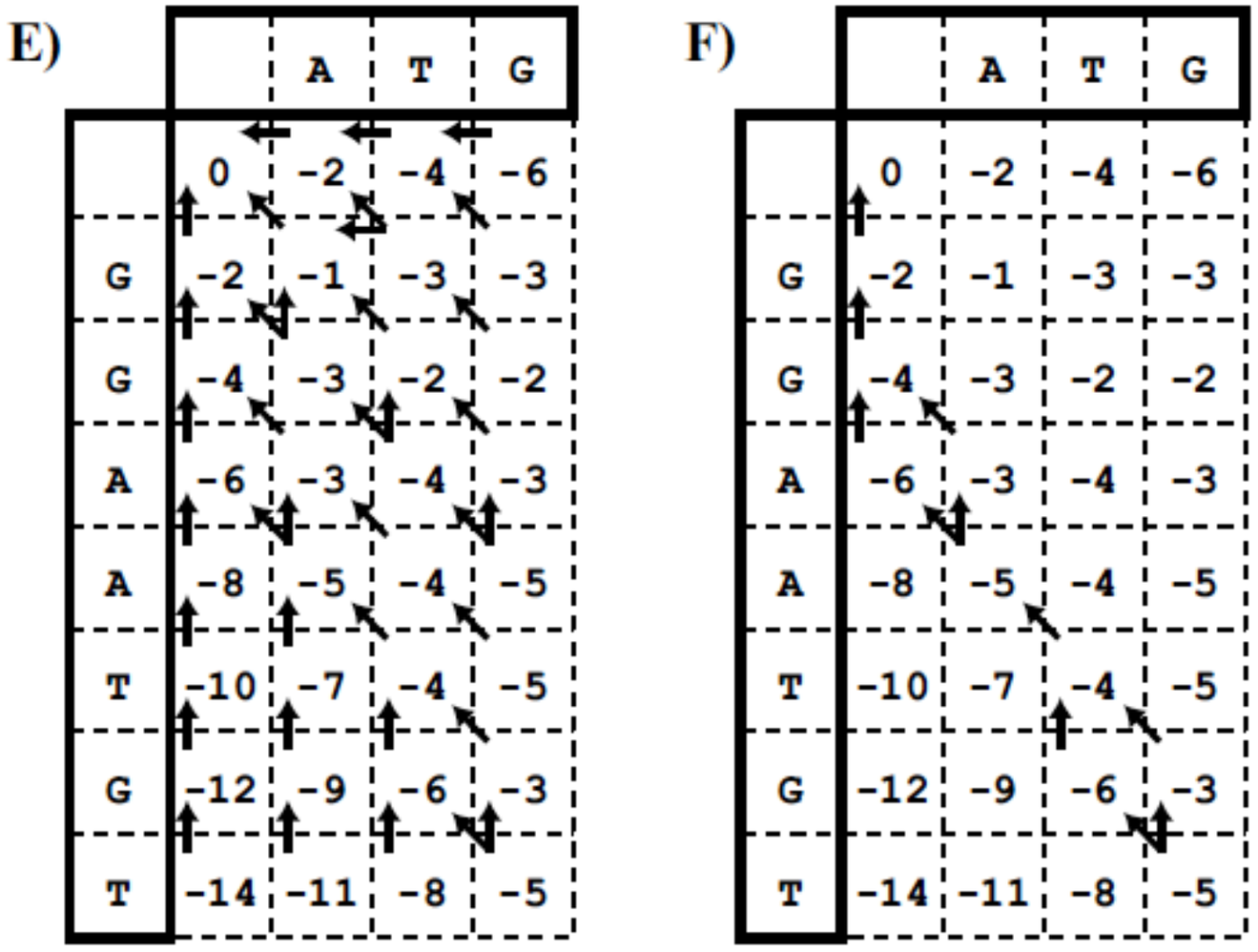
Alineamiento global

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.

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

Alineamiento 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.

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

Alineamiento global

```
GGAATGG    GGAATGG    GGAATGG    GGAATGG
---ATG-    ---AT-G    --A-TG-    --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.

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

Alineamiento global

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

```
AB--CDEF      ABCDEF      ABCDE--F
ABEDC--F      ABEDCF      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.

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

Alineamiento local

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

		A	T	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.

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

Alineamiento local

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

		A	T	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

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

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

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

¿Cómo determinar qué valores usar?

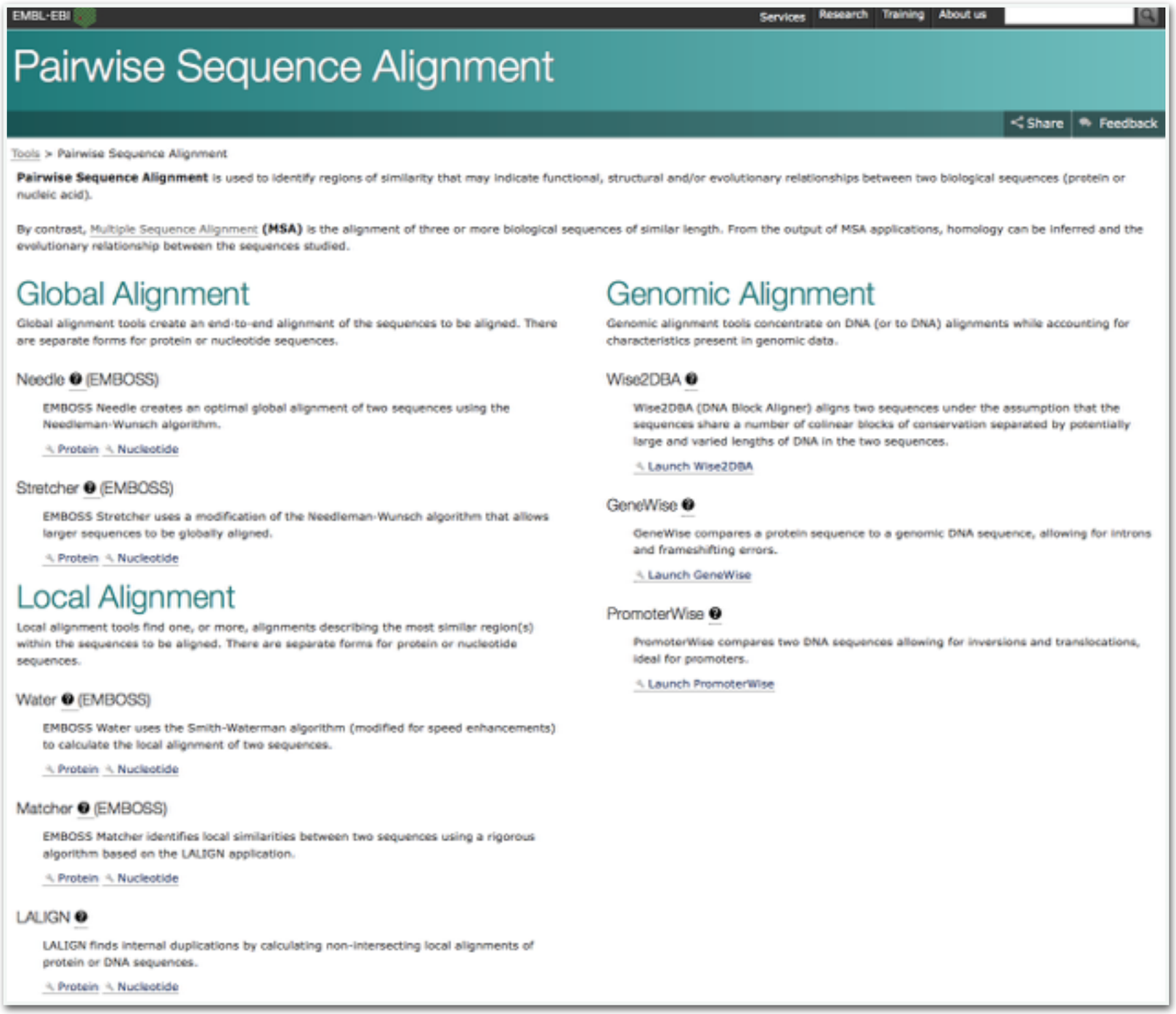
- La mayoría de la gente usa los valores “por defecto”, i.e., los que los autores pusieron
- proporción “mismatch”/“gap cost” lo más importante
- sustituciones/indels

```
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.

Sequence Alignment Methods, Models, Concepts, and Strategies Edited by Michael S. Rosenberg

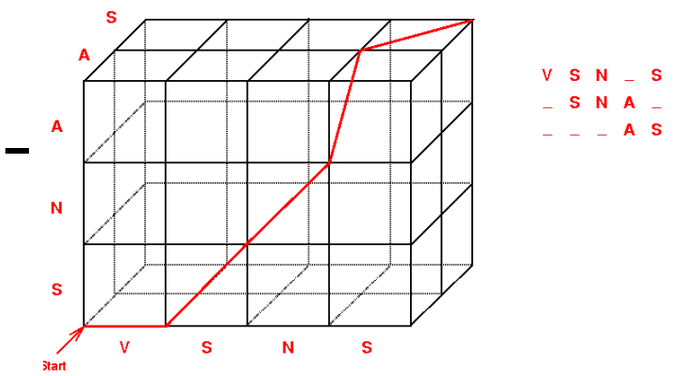
Buen lugar para empezar = http://www.ebi.ac.uk/Tools/psa/



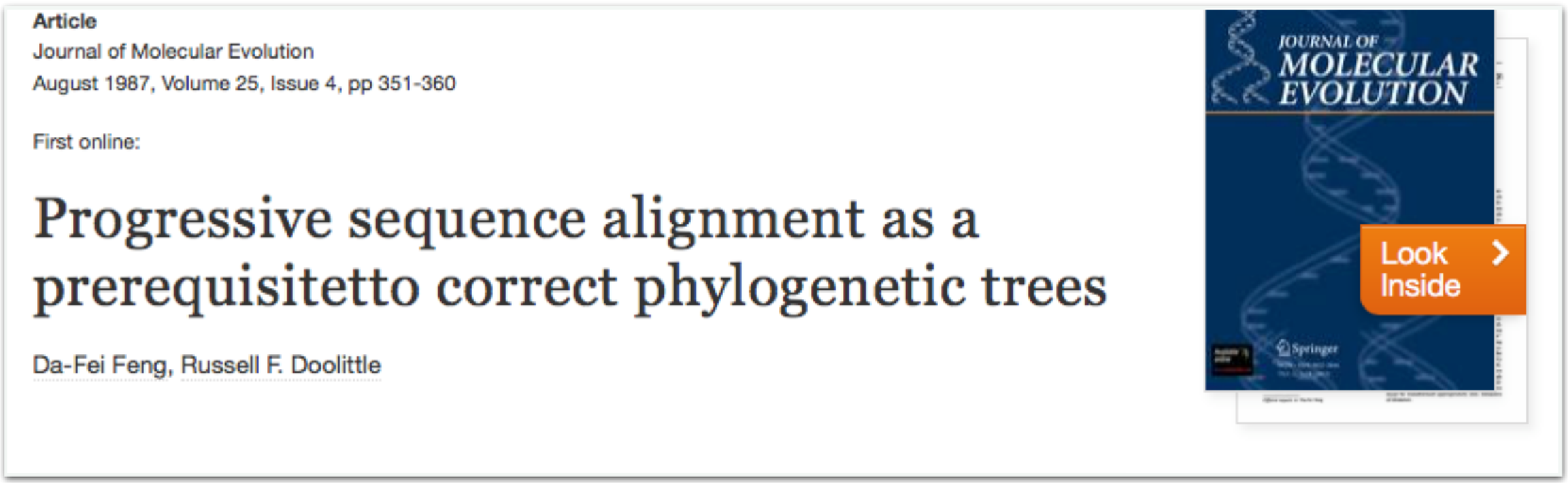
Alineamiento global múltiple (MSA)

Alineamiento múltiple

- Diferentes estrategias, e.g., matrices múlti-dimensionales (óptimo o exacto)



- 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 prerequisitetto correct phylogenetic trees

Da-Fei Feng, Russell F. Doolittle

Alineamiento múltiple: pasos

1. Calcular todos los alineamientos de pares

    1. para n secuencias, n x (n-1)/2 pares

2. Calcular dendrograma (guide tree*) usando algoritmo de clustering (UPGMA; Neighbor Joining)

3. Las secuencias más similares son alineadas primero de acuerdo al dendrograma

*"The quality of the guide-tree was found to affect MSA error levels only marginally." Landon and Graur, *Gene*, 2009

Alineamiento múltiple: pasos - versión gráfica

calcular alineamientos de a pares para todas las secuencias

S_1: PPGVKSDCAS
S_2: PADGVKDCAS
S_3: PPDGKSDS
S_4: GADGKDCCS
S_5: GADGKDCAS

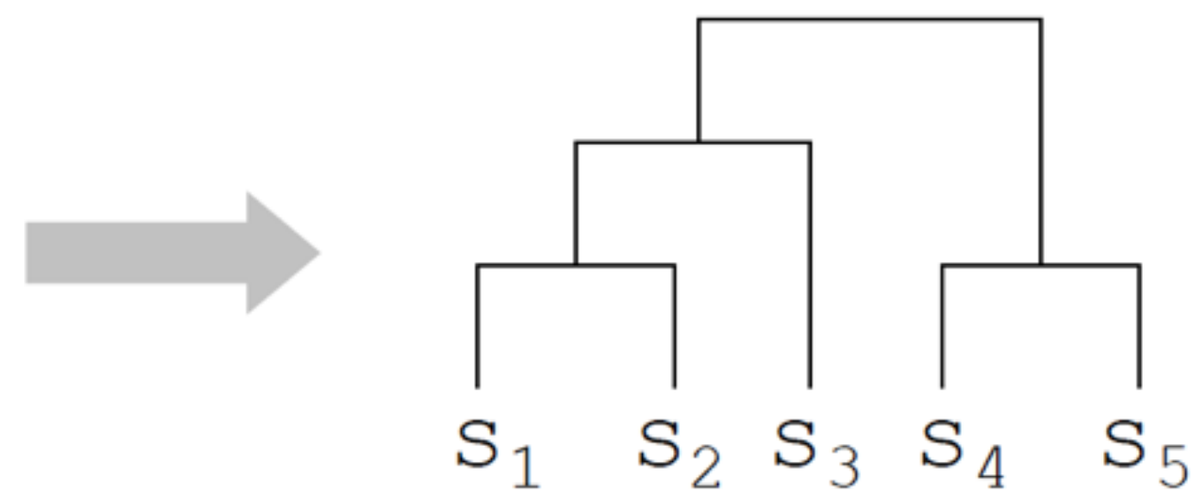
	S_1	S_2	S_3	S_4	S_5
S_1	0	0.111	0.25	0.555	0.444
S_2		0	0.375	0.222	0.111
S_3			0	0.5	0.5
S_4				0	0.111
S_5					0

http://www.comp.nus.edu.sg/~ksung/algo_in_bioinfo/slides/

Alineamiento múltiple: pasos - versión gráfica

crear un dendrograma guía (árbol guía)

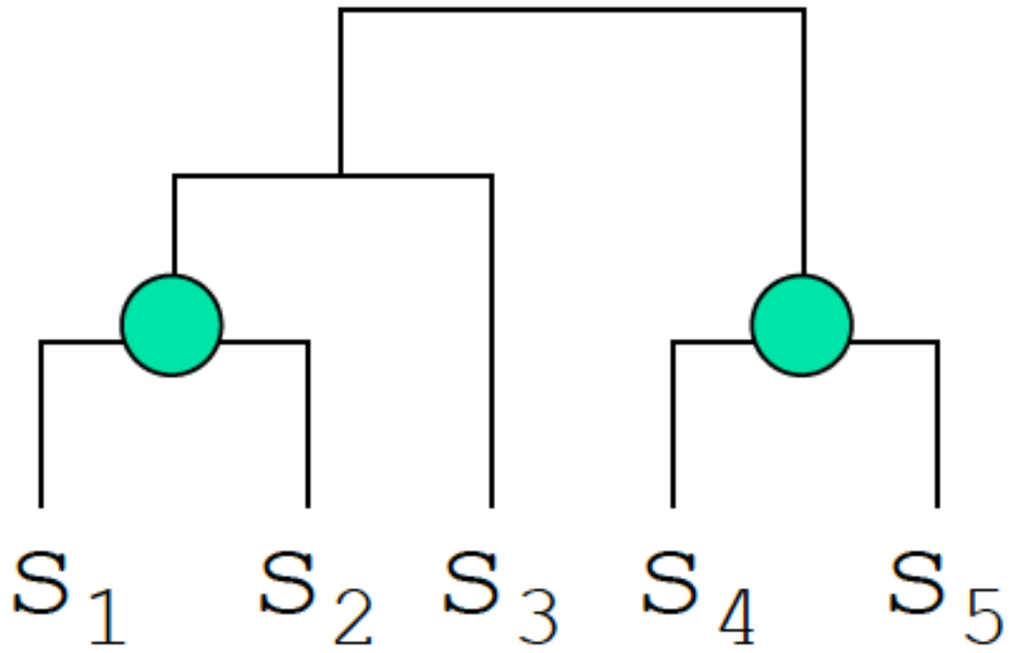
	S_1	S_2	S_3	S_4	S_5
S_1	0	0.111	0.25	0.555	0.444
S_2		0	0.375	0.222	0.111
S_3			0	0.5	0.5
S_4				0	0.111
S_5					0



http://www.comp.nus.edu.sg/~ksung/algo_in_bioinfo/slides/

Alineamiento múltiple: pasos - versión gráfica

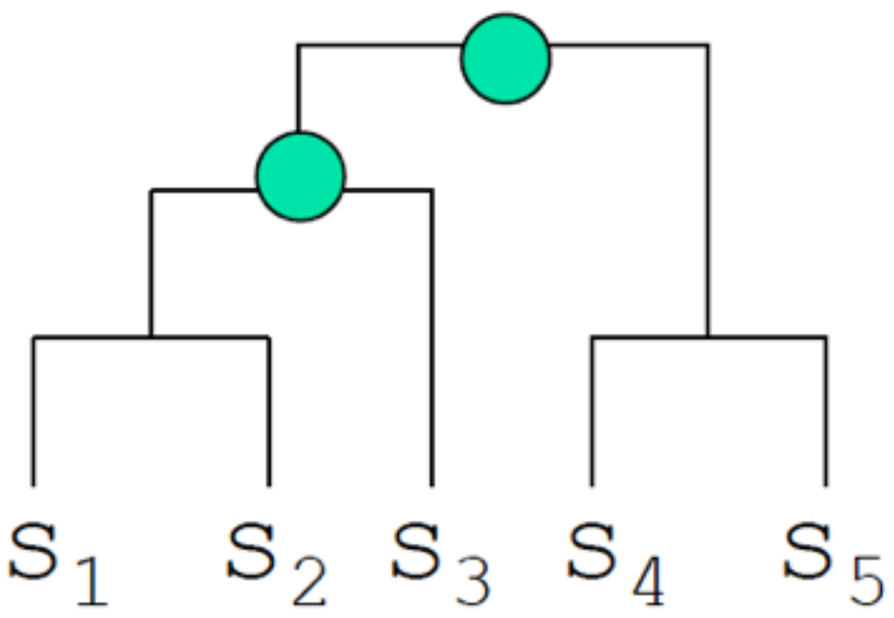
alineamos secuencias más similares primero



http://www.comp.nus.edu.sg/~ksung/algo_in_bioinfo/slides/

Alineamiento múltiple: pasos - versión gráfica

Alineamiento múltiple

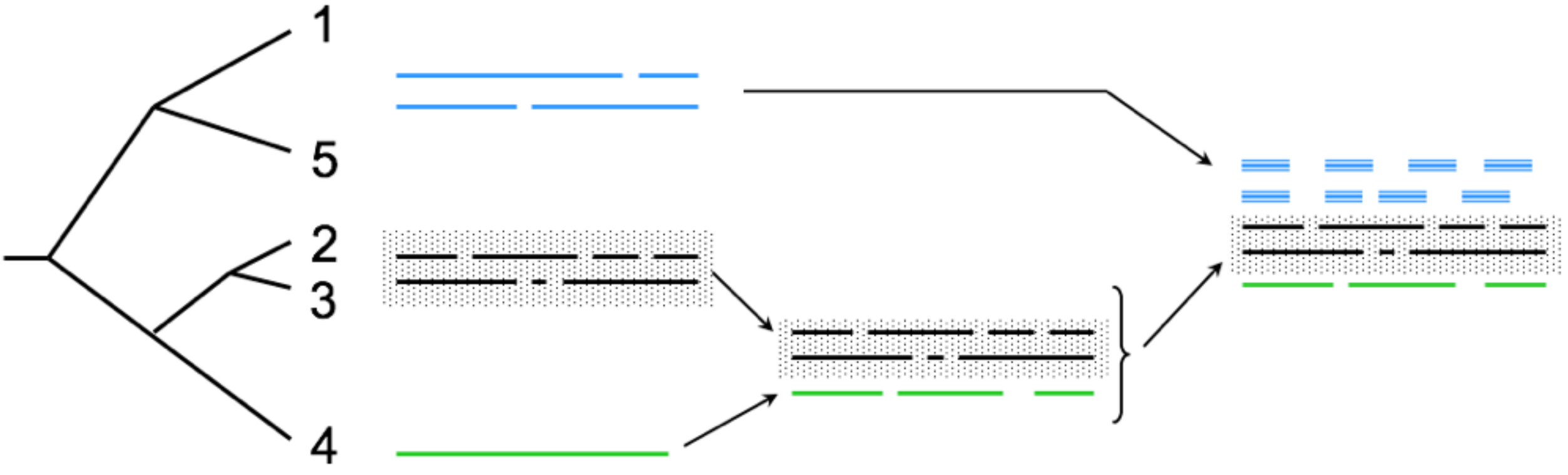



```
S1: P-PGVKSDCAS
S2: PADGVK-DCAS
S3: PPDG-KSD--S
S4: GADG-K-DCCS
S5: GADG-K-DCAS
```

http://www.comp.nus.edu.sg/~ksung/algo_in_bioinfo/slides/

Alineamiento múltiple: pasos - versión gráfica

Alineamiento múltiple



”The quality of the guide-tree was found to affect MSA error levels only marginally.” Landon and Graur, *Gene*, 2009

http://www.cs.uml.edu/~grinstei/91.510/msa_ppt.ppt

Alineamiento múltiple progresivo: desventajas

- Una vez que se abren gaps, éstos no pueden volver a cerrarse
- Errores en el alineamiento de las primeras secuencias son arrastrados y tienen efectos catastróficos sobre el MSA

Alineamiento múltiple: otros métodos

- métodos iterativos —> Muscle http://www.ebi.ac.uk/Tools/msa/muscle/. Dividir MSA en dos perfiles y realinear

- métodos de consenso —> M-Coffee http://www.tcoffee.org/Projects/mcoffee/

- métodos por segmentos —> DIALIGN http://dialign.gobics.de

- métodos estadísticos (modelos ocultos de Markov) —> HMMER http://hmmer.org

- algoritmos genéticos —> SAGA http://www.tcoffee.org/Projects/saga/

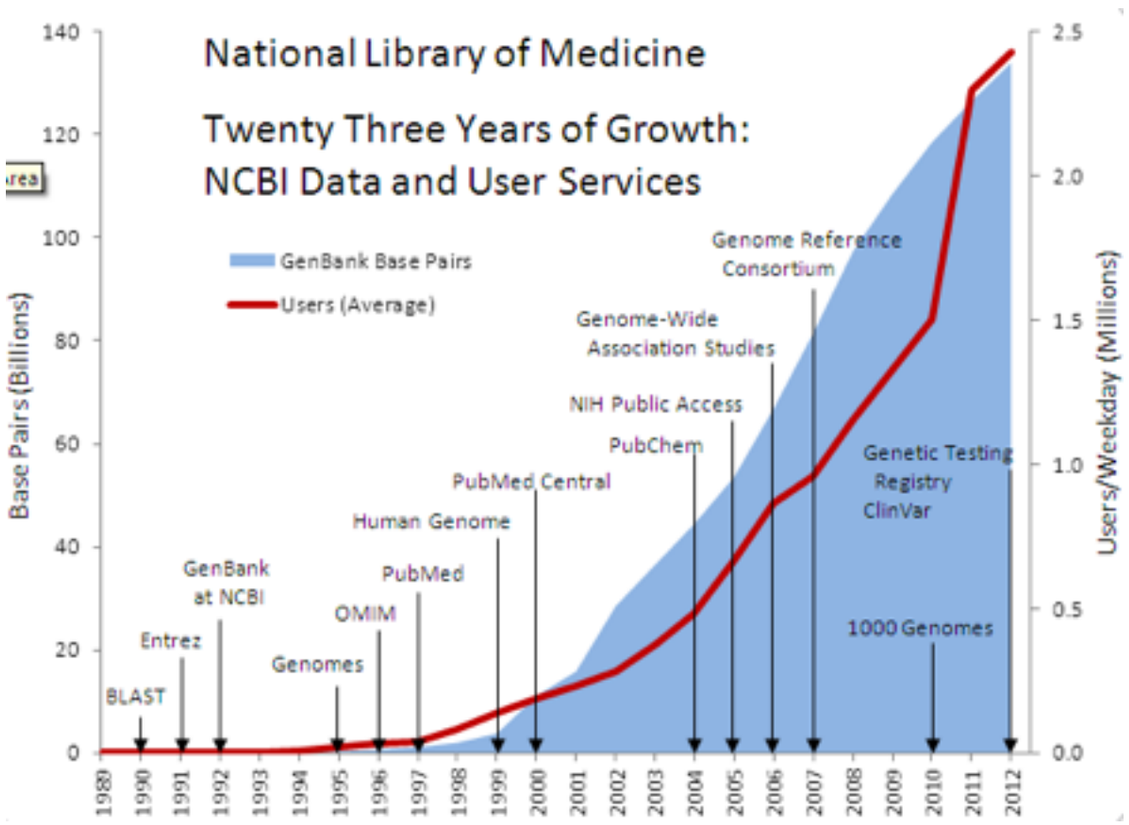
BLAST

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

¿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 ocurran por chance**

*pregunta de prueba

e-value

- **Representa el número de alineamientos diferentes con puntajes equivalentes o mejores que los que se esperan ocurran por chance**

- e-value bajo (10-5) puede ser indicativo de homología

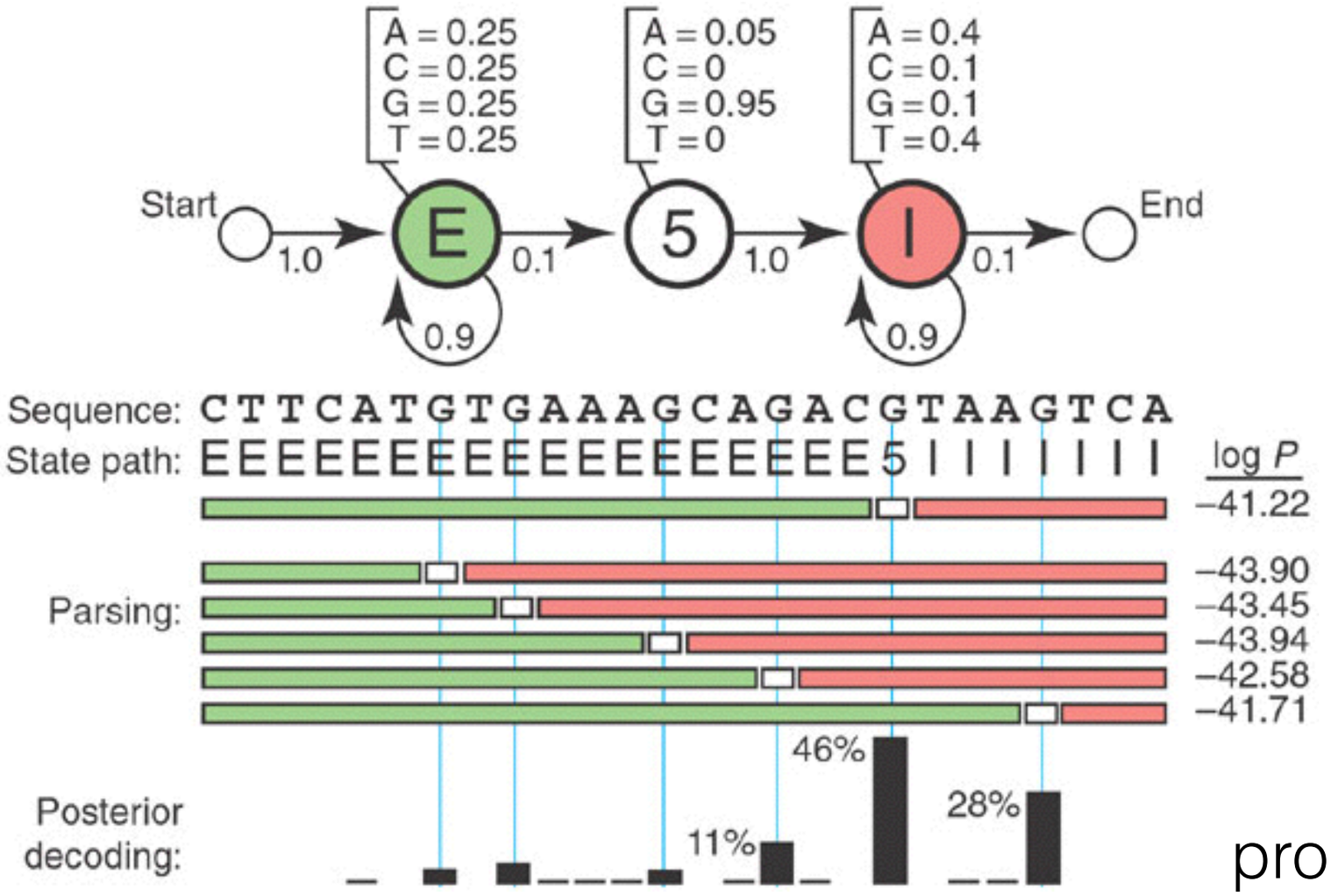
- Excepción: regiones de baja complejidad como repeticiones, pueden tener e-value bajo pero no ser homólogas

Modelos de Markov Ocultos (HMMs)

- En inglés “Hidden Markov Models”
- Son la caja de herramientas conceptual para generar un modelo probabilístico
- En vez de preguntarse cuáles son los valores de la matriz de costo para mi alineamiento, HMMs evaluar la probabilidad de que un número de combinaciones sea cierta

http://www.nature.com/nbt/journal/v22/n10/full/nbt1004-1315.html

Modelos de Markov Ocultos (HMMs)



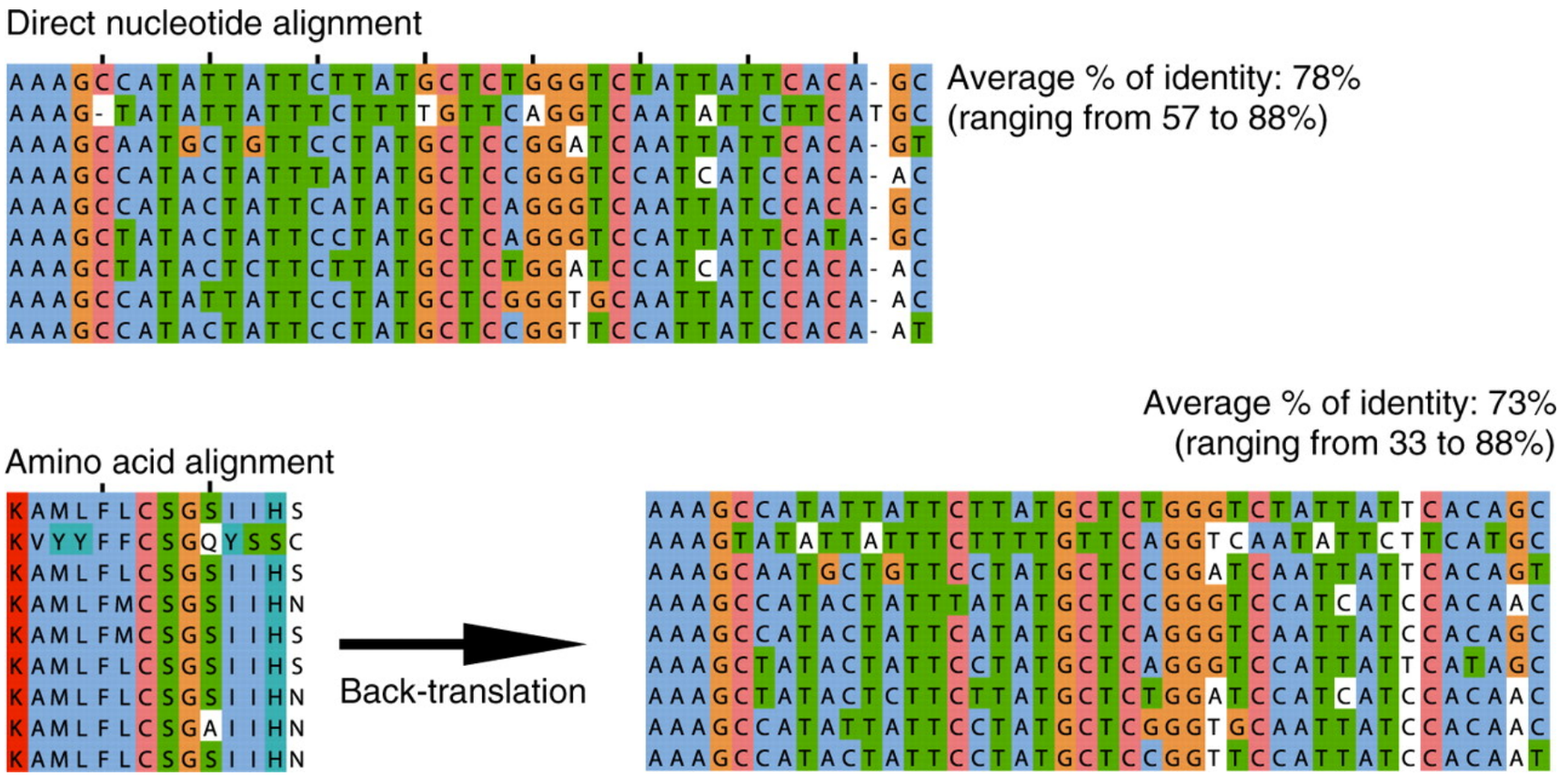
E exón
5 sitio de splicing
I intrón
“etiquetas”
secuencia observada

probabilidades de emisión
probabilidades de transición

Ejemplo para encontrar la mejor transición de exón a intrón

Estrategias para mejorar alineamientos

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

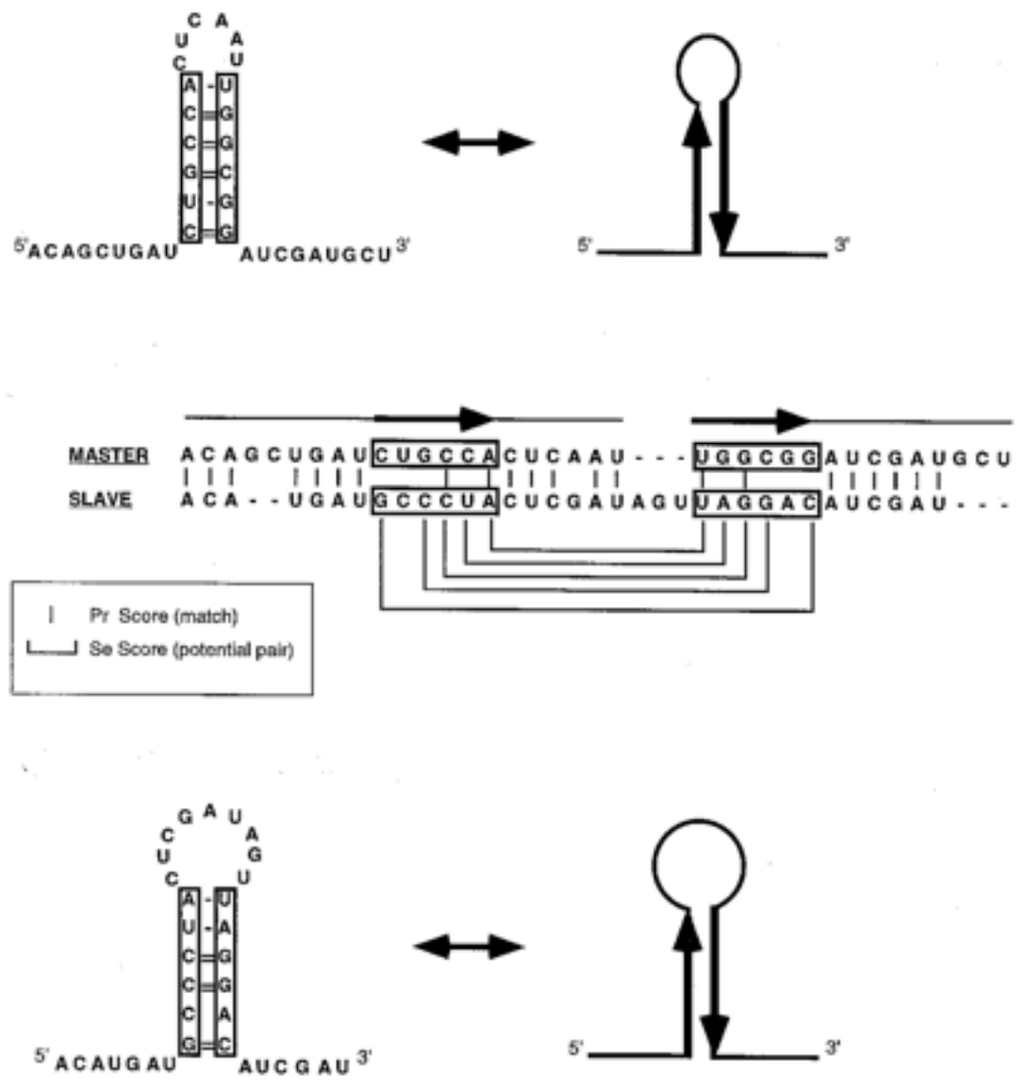


Federico Abascal et al. Nucl. Acids Res. 2010;38:W7-W13

TranslatorX; MACSE

Estrategias para mejorar alineamientos

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



http://tcoffee.org/Projects/raga/

Perfiles

- Una tabla que lista las frecuencias de cada nucleótido o aminoácido en cada posición de una secuencia

- Se calcula a partir de un MSA

- Ayuda a identificar una secuencia consenso

- Permite alinear nuevas secuencias al perfil —> búsqueda en base de datos / encontrar nuevas secuencias

- Al revés también, i.e., se usan perfiles para hacer MSAs

Position Specific Sequence Matrix (PSSM)

Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 4355–4358, July 1987
Biochemistry

Profile analysis: Detection of distantly related proteins

(amino acid/sequence comparison/protein structure/globin structure/immunoglobulin structure)

MICHAEL GRIBSKOV*, ANDREW D. MCLACHLAN†, AND DAVID EISENBERG*

*Molecular Biology Institute and Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90024; and †Medical Research Council, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England, United Kingdom

POS	PROBE	CONSENSUS		PROFILE																			
				A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L	V	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L L S P	L	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V V V V	V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K E A T	A	T	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A P L P	P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G G G G	G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S S Q E	D	E	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S S T P	S	P	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V L V A	V	A	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K R R S	R	S	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M L I I	I	I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S S T S	S	S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C C C C	C	C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K S Q R	K	R	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	A A G S	A	S	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T S D S	S	S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G G S Q	G	Q	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y F L S	F	S	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T T R L	T	L	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	F F . L	F	L	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S S . D	S	D	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S . . S	S	S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	. . . G	G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	. . . D	D	D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	. . . G	G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	. A G N	A	N	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y N Y T	Y	T	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E D D Y	D	Y	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L M A L	L	L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y N A W	N	W	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
.	.													.										
.	.													.										
.	.													.										
48	S G N S	S	S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S S N Y	S	Y	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

- Asignamos un puntaje a cada aminoácido en cada posición
- El puntaje del perfil para el aminoácido a en la posición p es:

$$M(p,a) = \sum_{b=1}^{20} f(p,b) \cdot s(a,b)$$

Donde

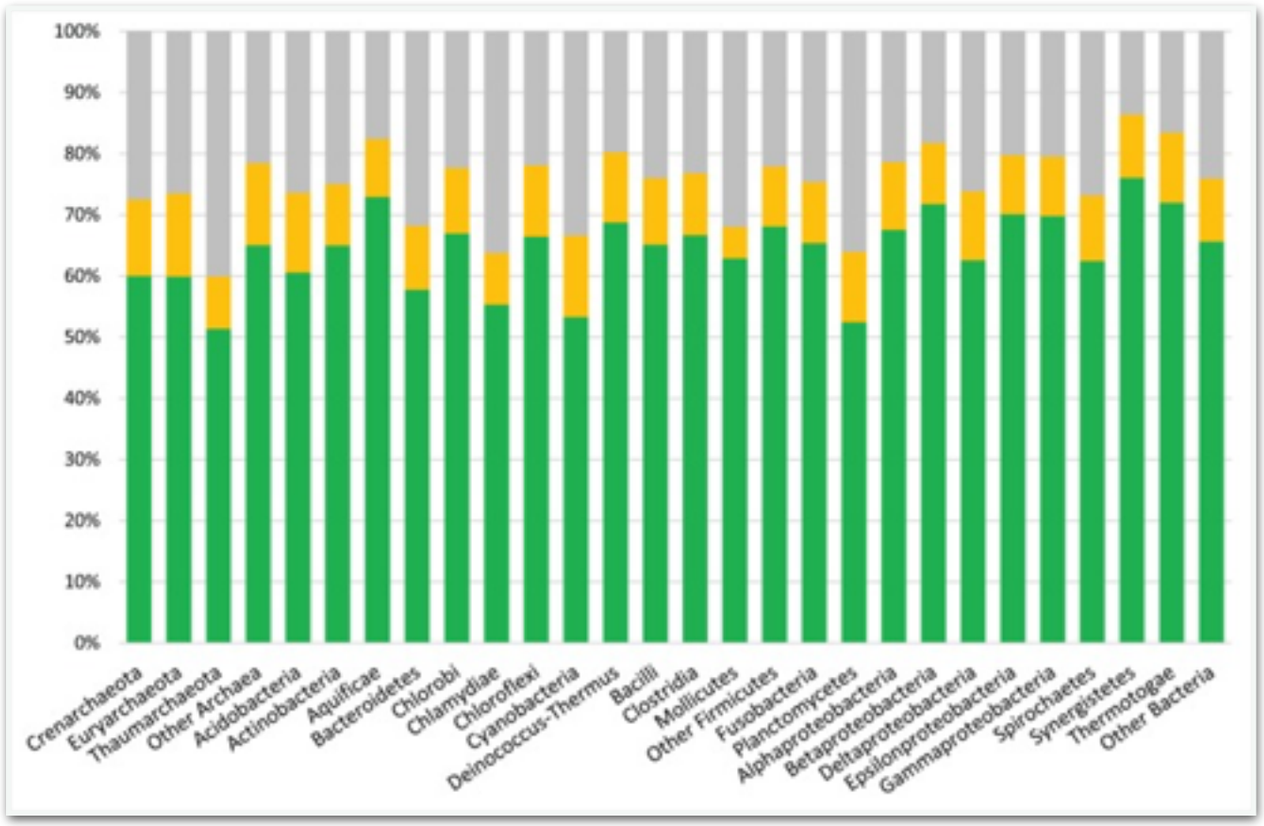
f(p,b) = frecuencia del aminoácido b en la posición p

s(a,b) es el puntaje de (a,b) (viene de e.g., BLOSUM or PAM)

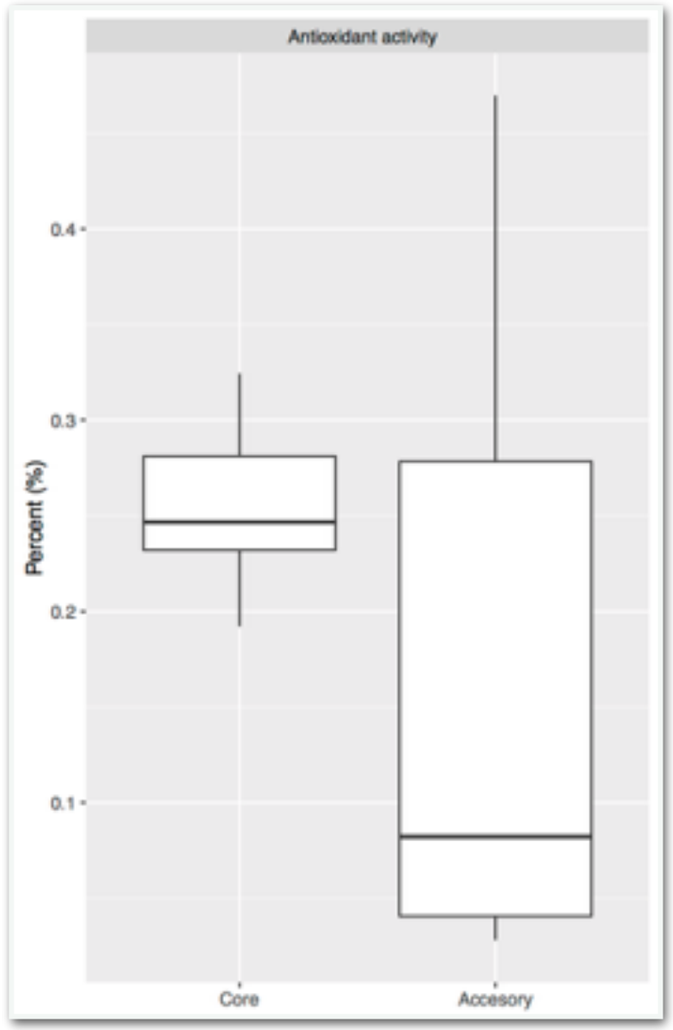
Ejemplo de PSSM —> COGs

- Clusters of Orthologous Groups, clasificación filogenética y funcional de grupos de proteínas ortólogas

- Secuencias —> PSI-BLAST —> PSSMs <— Query —> asignación de COG



COG coverage of various bacterial phyla. The columns represent the average fraction of proteins from the organisms in the given phylum that are not included in COGs (gray), assigned to the R or S categories in COGs (yellow) or assigned to other COG functional categories (green). For *Firmicutes* and *Proteobacteria*, coverage is shown at the class level. **62 a 711 genomas**



Galperin, M.Y., Makarova, K.S., Wolf, Y.I. and Koonin, E.V., 2014. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic acids research*, p.gku1223.

Recomendaciones

Program	Advantages	Cautions
CLUSTALW	Uses less memory than other programs	Less accurate or scalable than modern programs
DIALIGN	Attempts to distinguish between alignable and non-alignable regions	Less accurate than CLUSTALW on global benchmarks
MAFFT, MUSCLE	Faster and more accurate than CLUSTALW; good trade-off of accuracy and computational cost. Options to run even faster, with lower average accuracy, for high-throughput applications.	For very large data sets (say, more than 1000 sequences) select time- and memory-saving options
PROBCONS	Highest accuracy score on several benchmarks	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)
ProDA	Does not assume global alignability; allows repeated, shuffled and absent domains.	High computational cost and less accurate than CLUSTALW on global benchmarks
T-COFFEE	High accuracy and the ability to incorporate heterogeneous types of information	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)

Edgar, R.C. and Batzoglou, S., 2006. Multiple sequence alignment. *Current opinion in structural biology*, *16*(3), pp.368-373.

Recomendaciones

Input data	Recommendations
2–100 sequences of typical protein length (maximum around 10,000 residues) that are approximately globally alignable	Use PROBCONS, T-COFFEE, and MAFFT or MUSCLE, compare the results using ALTAVIST. Regions of agreement are more likely to be correct. For sequences with low percent identity, PROBCONS is generally the most accurate, but incorporating structure information (where available) via 3DCoffee (a variant of T-COFFEE) can be extremely helpful.
100–500 sequences that are approximately globally alignable	Use MUSCLE or one of the MAFFT scripts with default options. Comparison using ALTAVIST is possible, but the results are hard to interpret with larger numbers of sequences unless they are highly similar.
>500 sequences that are approximately globally alignable	Use MUSCLE with a faster option (we recommend maxiters-2) or one of the faster MAFFT scripts
Large numbers of alignments, high-throughput pipeline.	Use MUSCLE with faster options (e.g. maxiters-1 or maxiters-2) or one of the faster MAFFT scripts
2–100 sequences with conserved core regions surrounded by variable regions that are not alignable	Use DIALIGN
2–100 sequences with one or more common domains that may be shuffled, repeated or absent.	Use ProDA
A small number of unusually long sequences (say, >20,000 residues)	Use CLUSTALW. Other programs may run out of memory, causing an abort (e.g. a segmentation fault).

Edgar, R.C. and Batzoglou, S., 2006. Multiple sequence alignment. *Current opinion in structural biology*, *16*(3), pp.368-373.

Recomendaciones

“Garbage in - garbage out”

“It is not possible to carry out an accurate statistical analysis of inaccurate data”

"***All models are wrong, but some are useful***" George Box