

Búsquedas de secuencias en bases de datos

Heurísticas, Hits, significancia de los resultados

Fernán Agüero

Instituto de Investigaciones Biotecnológicas

Universidad Nacional de San Martín



Búsqueda de secuencias por similitud

- **Tenemos un método (algoritmo) que nos garantiza un alineamiento óptimo entre dos secuencias**
- **Tenemos un sistema de scoring complejo que refleja mejor nuestras ideas biológicas acerca de lo que es un alineamiento**
- **Cómo usaríamos estas herramientas para implementar una búsqueda por similitud contra una base de datos?**

Usemos la fuerza bruta

- Tenemos una base de datos con secuencias
- Tenemos una secuencia 'query' en la que estamos interesados
- Podemos encontrar secuencias similares al query en la base de datos?
- Tomar una por una las secuencias de la base de datos
- Calcular un alineamiento y su score
- Elegir los mejores alineamientos en base al score
- Finalmente usar nuestro criterio y evaluar si la/s secuencia/s encontradas son lo suficientemente similares

Heurísticas para reducir espacio de búsqueda

- **Hay dos espacios de búsqueda reconocibles:**
 - El espacio de todas las secuencias de la base de datos
 - El espacio de todos los alineamientos posibles entre dos secuencias
- **Las búsquedas de secuencias por similitud son “exactas” si recorren completamente ambos espacios**
 - Ej: Smith-Waterman sobre toda la base de datos
- **A continuación vamos a introducir heurísticas para reducir estos espacios de búsqueda**
 - Estrategias de hashing para filtrar la base de datos
 - Distintas heurísticas para reducir el espacio de alineamientos posibles que se explora efectivamente

Búsquedas en bases de datos

**Compara una
secuencia (query)
contra una base de
datos de secuencias**

Una búsqueda
típica tiene 4
elementos básicos.

```
> fasta myquery swissprot -ktup 2
```

Programa query Base de datos Parámetros opcionales

Búsqueda en bases de datos

Con el crecimiento exponencial de las bases de datos las búsquedas son cada vez más lentas ...

```
> fasta myquery swissprot -ktup 2  
  
searching .....
```

Database searching

La lista de hits provee los 'títulos' y scores de las secuencias que fueron seleccionadas por la secuencia 'query'.

> fasta myquery swissprot -ktup 2

```
The best scores are:
                                initn initl opt  z-sc E(77110)
gi|1706794|sp|P49789|FHIT_HUMAN BIS(5'-ADENOSYL)- 996 996 996 1262.1 0
gi|1703339|sp|P49776|APH1_SCHPO BIS(5'-NUCLEOSYL) 412 382 395 507.6 1.4e-21
gi|1723425|sp|P49775|HNT2_YEAST HIT FAMILY PROTEI 238 133 316 407.4 5.4e-16
gi|3915958|sp|Q58276|Y866_METJA HYPOTHETICAL HIT- 153 98 190 253.1 2.1e-07
gi|3916020|sp|Q11066|YHIT_MYCTU HYPOTHETICAL 15.7 163 163 184 244.8 6.1e-07
gi|3023940|sp|O07513|HIT_BACSU HIT PROTEIN 164 164 170 227.2 5.8e-06
gi|2506515|sp|Q04344|HNT1_YEAST HIT FAMILY PROTEI 130 91 157 210.3 5.1e-05
gi|2495235|sp|P75504|YHIT_MYCPN HYPOTHETICAL 16.1 125 125 148 199.7 0.0002
gi|418447|sp|P32084|YHIT_SYNP7 HYPOTHETICAL 12.4 42 42 140 191.3 0.00058
gi|3025190|sp|P94252|YHIT_BORBU HYPOTHETICAL 15.9 128 73 139 188.7 0.00082
gi|1351828|sp|P47378|YHIT_MYCGE HYPOTHETICAL HIT- 76 76 133 181.0 0.0022
gi|418446|sp|P32083|YHIT_MYCHR HYPOTHETICAL 13.1 27 27 119 165.2 0.017
gi|1708543|sp|P49773|IPK1_HUMAN HINT PROTEIN (PRO 66 66 118 163.0 0.022
gi|2495231|sp|P70349|IPK1_MOUSE HINT PROTEIN (PRO 65 65 116 160.5 0.03
gi|1724020|sp|P49774|YHIT_MYCLE HYPOTHETICAL HIT- 52 52 117 160.3 0.031
gi|1170581|sp|P16436|IPK1_BOVIN HINT PROTEIN (PRO 66 66 115 159.3 0.035
gi|2495232|sp|P80912|IPK1_RABIT HINT PROTEIN (PRO 66 66 112 155.5 0.057
gi|1177047|sp|P42856|ZB14_MAIZE 14 KD ZINC-BINDIN 73 73 112 155.4 0.058
gi|1177046|sp|P42855|ZB14_BRAJU 14 KD ZINC-BINDIN 76 76 110 153.8 0.072
gi|1169825|sp|P31764|GAL7_HAEIN GALACTOSE-1-PHOSP 58 58 104 138.5 0.51
gi|113999|sp|P16550|APA1_YEAST 5',5'''-P-1,P-4-TE 47 47 103 137.8 0.56
gi|1351948|sp|P49348|APA2_KLULA 5',5'''-P-1,P-4-T 63 63 98 131.3 1.3
gi|123331|sp|P23228|HMCS_CHICK HYDROXYMETHYLGLUTA 58 58 99 129.4 1.6
gi|1170899|sp|P06994|MDH_ECOLI MALATE DEHYDROGENA 70 48 91 122.9 3.7
gi|3915666|sp|Q10798|DXR_MYCTU 1-DEOXY-D-XYLULOSE 75 50 92 121.9 4.3
gi|124341|sp|P05113|IL5_HUMAN INTERLEUKIN-5 PRECU 36 36 85 121.3 4.7
gi|1170538|sp|P46685|IL5_CERTO INTERLEUKIN-5 PREC 36 36 84 120.0 5.5
gi|121369|sp|P15124|GLNA_METCA GLUTAMINE SYNTHETA 45 45 90 118.9 6.3
gi|2506868|sp|P33937|NAP_A_ECOLI PERIPLASMIC NITRA 48 48 92 117.4 7.6
gi|119377|sp|P10403|ENV1_DROME RETROVIRUS-RELATED 59 59 89 117.0 8
gi|1351041|sp|P48415|SC16_YEAST MULTIDOMAIN VESIC 48 48 97 117.0 8
gi|4033418|sp|O67501|IPYR_AQUAE INORGANIC PYROPHO 38 38 83 116.8 8.3
```

El detalle de los alineamientos se muestra más abajo

```
>>gi|1703339|sp|P49776|APH1_SCHPO BIS(5'-NUCLEOSYL)-TETR (182 aa)
initn: 412  initl: 382  opt: 395  z-score: 507.6  E(): 1.4e-21
Smith-Waterman score: 395;      52.3% identity in 109 aa overlap
```

```

      60      70      80      90     100     110
gi|170 QTTQRVGTVVEKHFHGTSLTFSMQDGPEAGQTVKHVHVHVLPRKAGDFHRNDSIYEELQK
      . . . . :  : . . . : . . . . . : . . . . : . . . . . : . . : . : : X. :
gi|170 TSVRKVQQVIEKVFSSASASNIGIQDGVDAQTVPHVHVHIIIPRKKADFSENDLVYSELEK
      70      80      90     100     110     120

```

```

      120      130      140
gi|170 HDKEDFPASWRSEEEEMAAEAAALRVYFQ
      ..
gi|170 NEGNLASLYLTGNERYAGDERPPTSMRQAIPKDEDRKPRTL EEMEKEAQWLKGYFSEEQE
      130      140      150      160      170      180

```

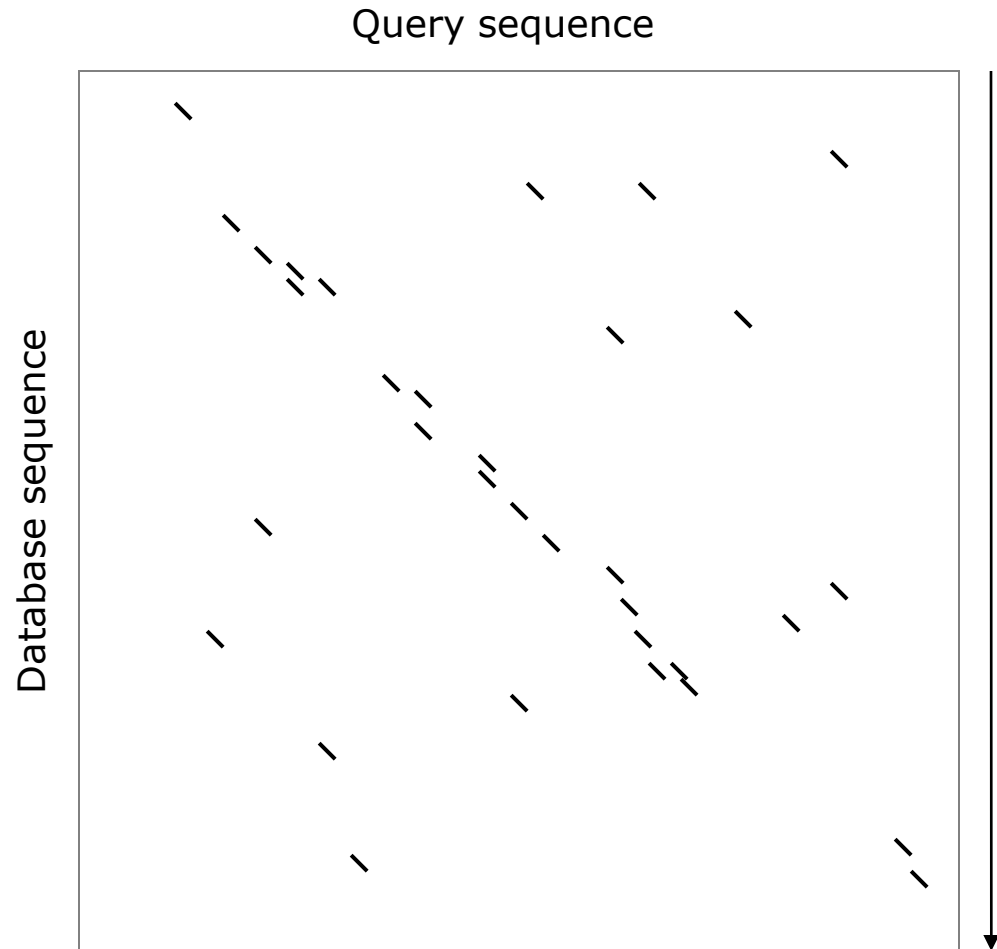
```
>>gi|1723425|sp|P49775|HNT2_YEAST HIT FAMILY PROTEIN 2 (217 aa)
initn: 238  initl: 133  opt: 316 z-score: 407.4 E(): 5.4e-16
Smith-Waterman score: 316;      37.4% identity in 131 aa overlap
```

gi|170 MSFRFGQHLIKPSVVF¹⁰FLKTELSFALVNRKPVVPGHVLVCPLRP²⁰VER³⁰FER⁴⁰

Búsquedas en bases de datos: hashing methods

La búsqueda más simple es un gran ejemplo de dynamic programming. Para una secuencia query de **N** letras, contra una base de datos de **M** letras, se requieren **$M \times N$** comparaciones.

Cómo reducir este espacio de búsqueda?



Hashing methods

Hashing es un método común para acelerar búsquedas en bases de datos.

Compilar un “diccionario” de palabras a partir de la secuencia ‘query’. Armar un índice con todas las palabras.

Longitud de la secuencia:
19

Cantidad de palabras:
17

MLIIKRDELVISWASHERE

MLI
LII
IIK
IKR
KRD
RDE
DEL
ELV
LVI
VIS
ISW
SWA
WAS
ASH
SHE
HER
ERE

query
sequence

Todas las palabras
posibles de
longitud **ktup**

ktup = 3

Hashing methods

Construir el diccionario de palabras para la secuencia 'query' requiere $N-2$ operaciones.

La base de datos contiene $M-2$ palabras, con un límite máximo de 20^{ktup} palabras (proteínas = 20 aminoácidos o letras posibles)

Para $ktup=3$ el número total (máximo) de palabras es $20^3 = 8000$

Esta operación de búsqueda es muy eficiente computacionalmente

MLIIKRDELVISWASHERE

**MLI
LII
IIK
IKR
KRD
RDE
DEL
ELV
LVI
VIS
ISW
SWA
WAS
ASH
SHE
HER
ERE**

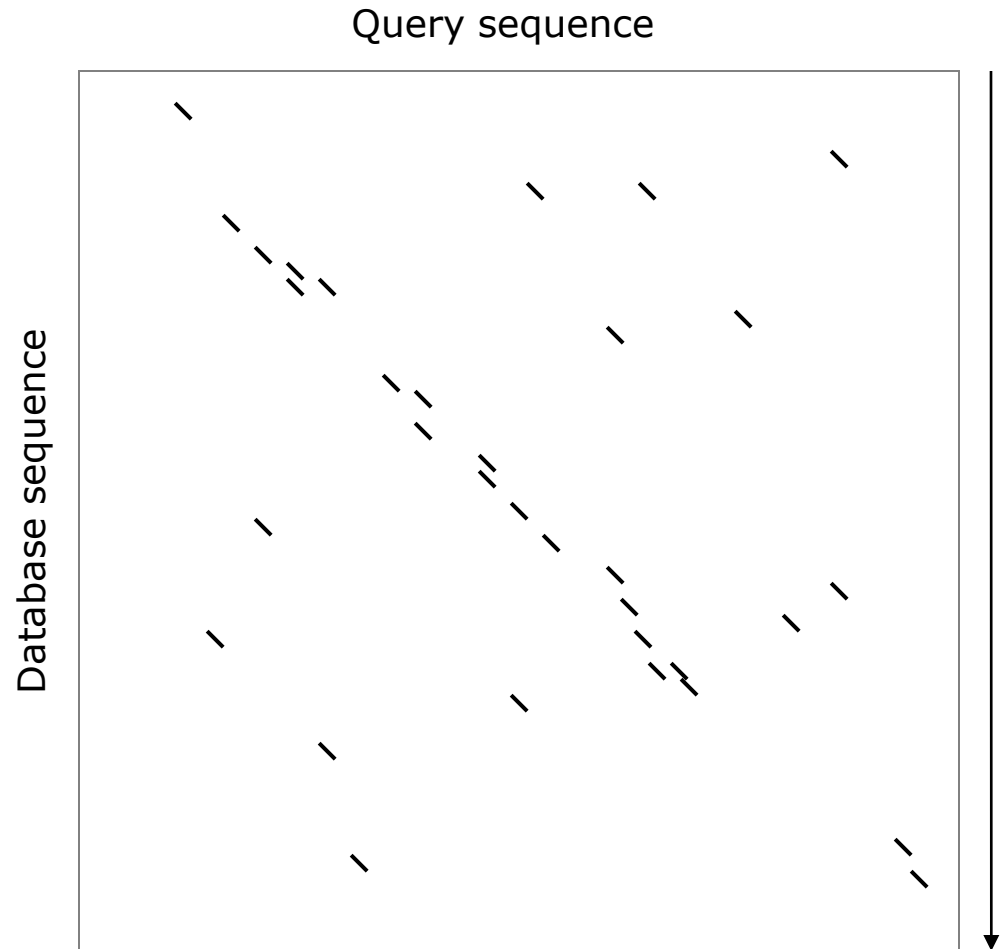
query
sequence

all overlapping
words of size 3

Hashing methods

**Scan the database,
looking up words in
the dictionary**

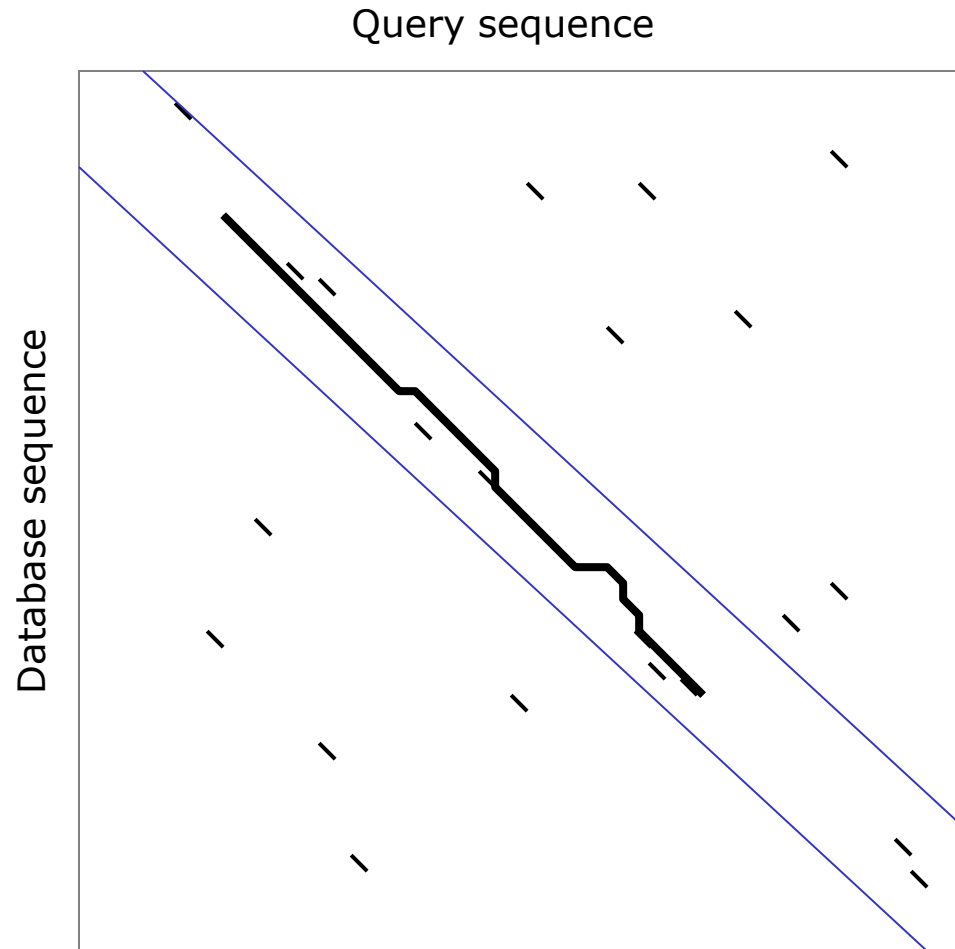
Use word hits to
determine where to search
for alignments
fills the dynamic
programming matrix
in $(N-2)+(M-2)$ operations
instead
of $M \times N$.



Hashing methods

**Scan the database,
looking up words in
the dictionary**

Use word hits to
determine where to search
for alignments

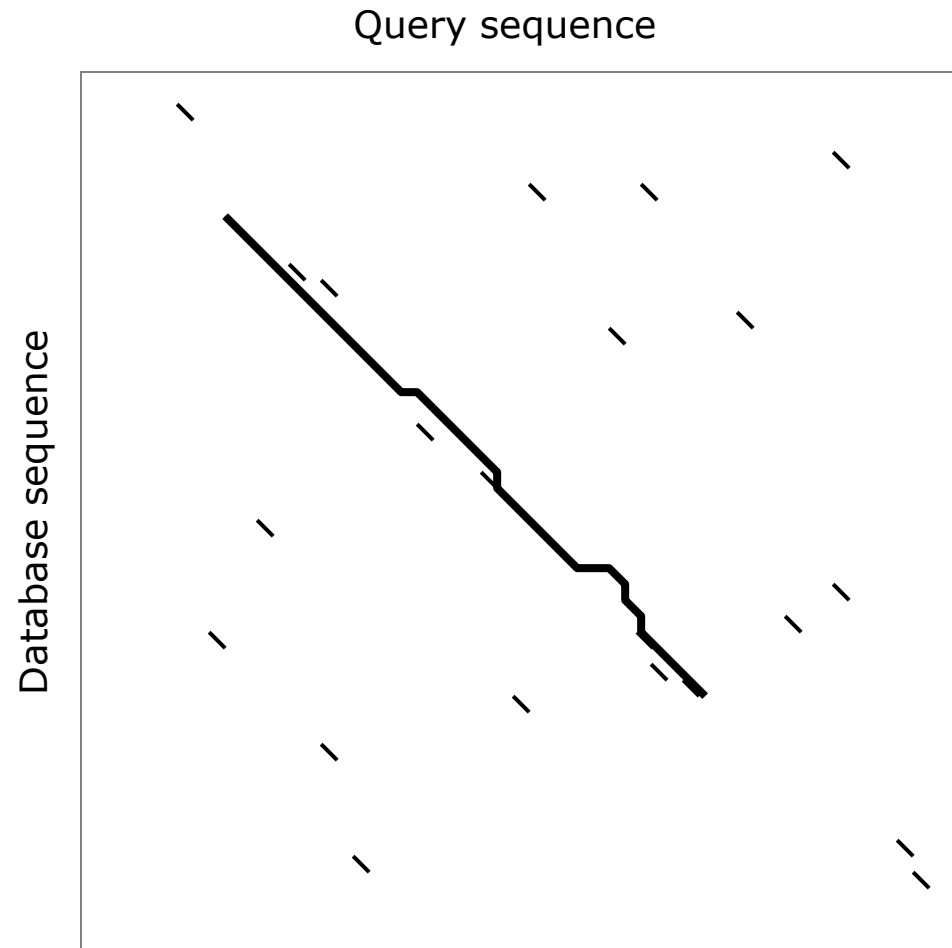


FASTA searches in a band

Hashing methods

**Scan the database,
looking up words in
the dictionary**

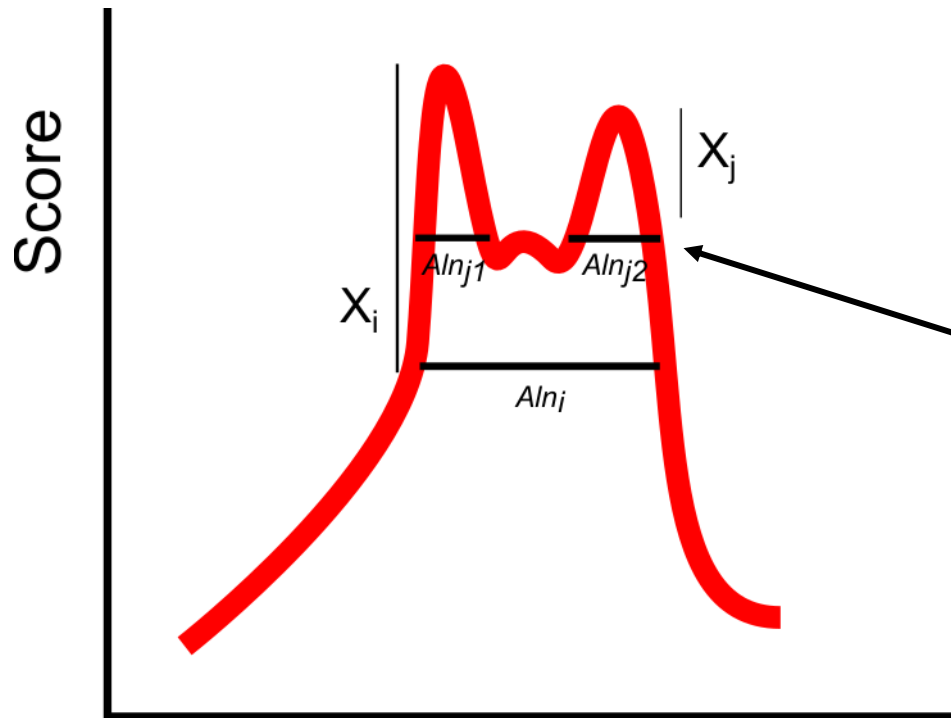
Use word hits to
determine where to search
for alignments



BLAST extends from word hits

BLAST: varios HSPs

un HSP es un “*high scoring pair*”

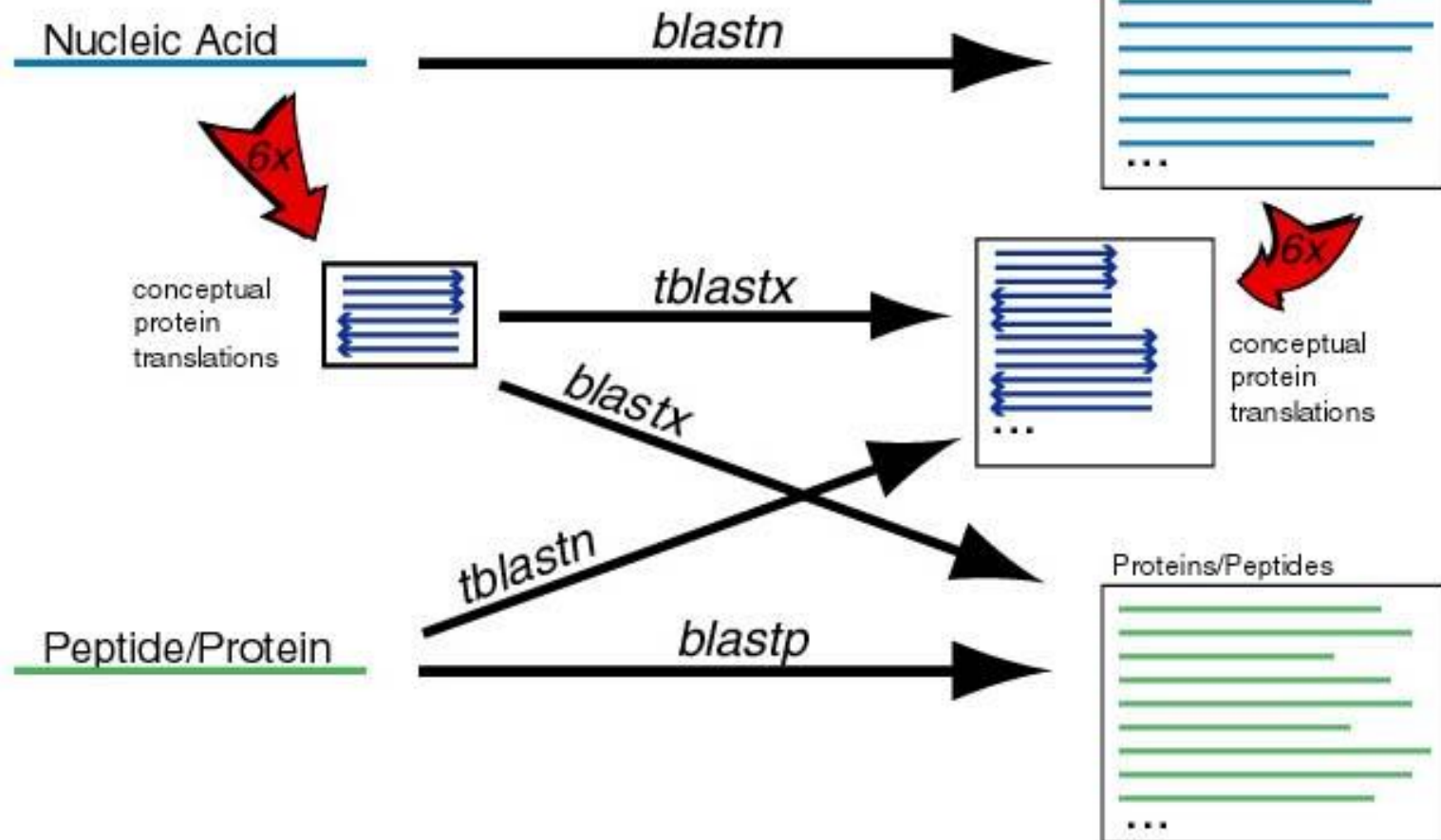


BLAST intenta extender el HSP, siempre que la caída del score sea menos que X (bits). Si lo logra, se repite con el próximo pico.

BLAST: algoritmos

QUERY
SEQUENCE

DATABASE



FASTA: algoritmos

- **FASTA**
 - protein-protein, DNA-DNA
- **fastx, fasty**
 - translated query, protein database
 - Permite frameshifts sólo entre codones (fastx) o dentro de un codón (fasty)
- **Ssearch**
 - Una implementación rigurosa del algoritmo de Smith-Waterman (sin heurísticas)
- **Prss**
 - Evalúa el significado de un alineamiento por permutación de una secuencia
- **Tfastx, tfasty**
 - Protein sequence vs DNA database

Evaluando alineamientos

- **Qué hacemos cuando estamos comparando dos secuencias que no son claramente similares, pero que muestran un alineamiento prometedor?**
- **Necesitamos un test de significancia**
- **Tenemos que responder a la pregunta:**
 - **Cuál es la probabilidad de que un alineamiento similar (con un score similar) ocurra entre proteínas no relacionadas?**

- **Generar secuencias al azar de la misma longitud y composición que la secuencia query y alinearlas**
 - Karlin & Altschul (1990); Altschul et al (1994); Altschul & Gish (1996)
- **Analizar la distribución de scores que se obtiene**

The Gumbel/Extreme value distribution

- In a database search (BLAST/FASTA) the alignment scores **do not** follow a normal/Gaussian distribution!

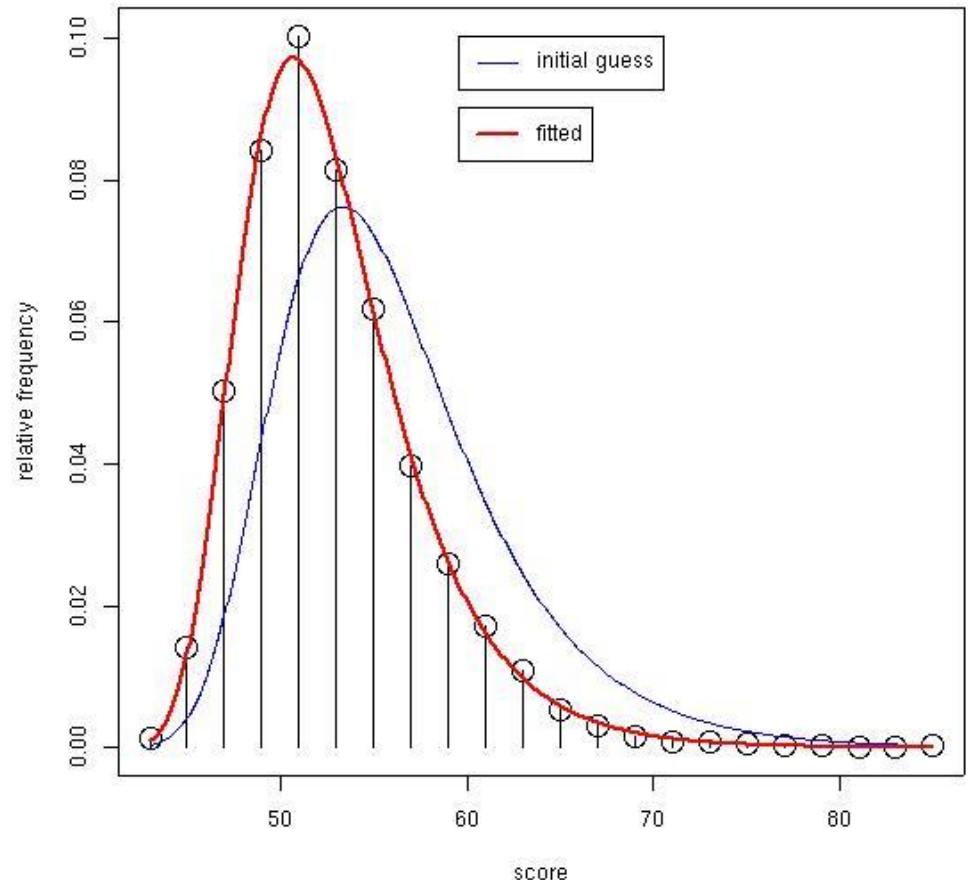
$$E = K m n e^{-\lambda S}$$

E is the number of alignments with a score = S

m,n: length of the sequences

K,λ: estimated parameters estimated (depend on the scoring matrix and the size of the database)

Extreme Value Distribution, Empirical method



E-value

Los hits pueden ser ordenados de acuerdo a su E-value o a su Score.

El E-value – más conocido como **EXPECT** value – es una función del score, el tamaño de la base de datos y de la longitud de la secuencia 'query'.

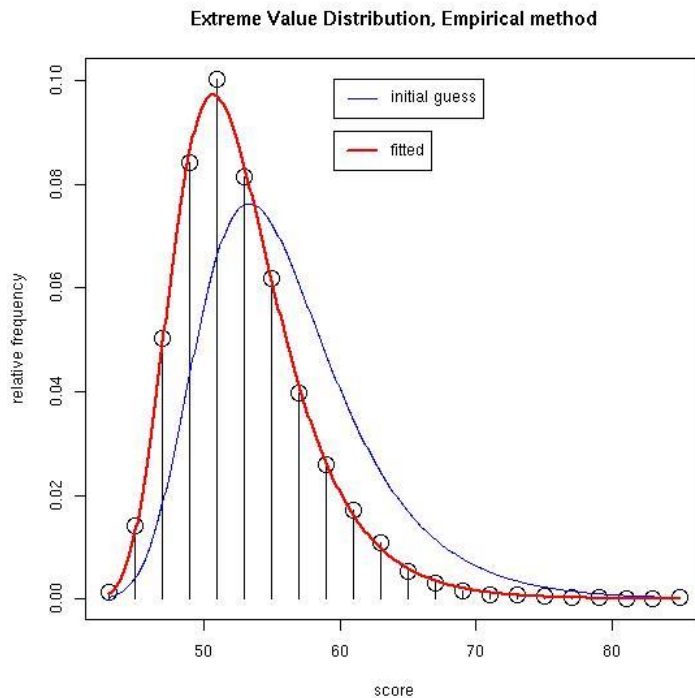
E-value: Número de alineamientos con un score = S que se espera encontrar si la base de datos es una colección de letras al azar.

Ejemplo: En el caso de un score=1 (un match o identidad) debería haber un número enorme de alineamientos. Uno espera encontrar menos alineamientos con un score de 5, 10, etc. Eventualmente, cuando el score es lo suficientemente alto, uno espera encontrar un número insignificante de alineamientos que sean debidos al azar.

Valores de E-value menores que $1e-6$ ($1 * 10^{-6}$) son generalmente muy buenos para proteínas, mientras que $E < 1e-2$ puede considerarse significativo. Es posible que un hit cuyo $E > 1$ sea biológicamente importante, aunque es necesario analizarlo más detalladamente para confirmarlo.

Observed vs expected

- Si la base de datos es suficientemente grande y contiene mayoritariamente secuencias no relacionadas la distribución de scores **observados** debería coincidir bastante con la distribución de scores **esperados** por azar (Pearson 1998)



Tamaño de la base de datos

$$E(S > x) = p(S > x) \text{ D}$$

- El número de alineamientos con un score $> S$ se incrementa linealmente con el tamaño de la base de datos
- \Rightarrow una secuencia (un alineamiento con un score S) encontrada en una búsqueda contra un genoma bacteriano con 1000-5000 secuencias va a ser 50-250 veces más significativa que un alineamiento con exactamente el mismo score en una base de datos como OWL (250,000 secuencias)
- Sin embargo, vimos que la base de datos tiene que ser **suficientemente grande** como para poder estimar P y E
- \Rightarrow Compromiso

Tamaño de la base de datos: un ejemplo

- **Objetivo:** encontrar el homólogo en *E. coli* de la DAHP synthase de *B. subtilis*
- ***E. coli* proteome**
 - **kdsA, $E(4283) < 0.00015$**
- **Swissprot**
 - **kdsA, $E(74417) < 0.0017$**
- **OWL**
 - **kdsA, $E(260784) < 0.0085$**
- **El mismo alineamiento, con el mismo score es 50 veces más significativo en la base de datos más chica.**

Identificar homólogos con eficiencia

- **Buscar en bases de datos pequeñas primero**
- **Repetir la búsqueda en una base de datos pequeña con un algoritmo más sensible (fasta3 con ktup 1 o ssearch)**
- **Si no hay hits significativos, buscar bases de datos más grandes, como nr (GenPept, TrEMBL)**

Límites de la estadística

- **En ciertos casos, la estadística de los alineamientos falla**
 - Lo que falla son las suposiciones que hicimos para llegar al modelo estadístico que describe - en este caso - la distribución de scores entre secuencias no relacionadas
- **En general se obtienen estimaciones incorrectas de E cuando**
 - Se usan penalidades de gap incorrectas
 - Existen regiones de baja complejidad en la secuencia query

Evaluando la estadística

```
opt      E()
< 20    13    0:=
22      0     0:
24      0     0:
26      0     0:
28      1     3:*
30     11    19:*
32     46    75:===*
34    242   204:=====+
36    493   419:=====+====
38    788   692:=====+=====
40   1055   965:=====+=====
42   1275  1180:=====+=====
44   1299  1302:=====+=====
46   1251  1326:=====+=====
48   1186  1269:=====+=====
50   1077  1158:=====+=====
52    907  1018:=====+=====
54    849   870:=====+=====
56    714   727:=====+=====
58    570   596:=====+=====
60    456   483:=====+=====
62    393   387:=====+=====
64    313   308:=====+=====
66    268   243:=====+=====
68    219   192:=====+=====
70    191   150:=====+=====
72    127   117:=====+=====
74     93    91:=====+=====
76     91    71:=====+=====
78     44    55:=====+=====
80     33    43:=====+=====
82     22    33:=====+=====
84     32    26:=====+=====
86     19    20:=====+=====
88     19    16:=====+=====
90      8    12:=====+=====
92      8     9:=====+=====
94      5     7:=====+=====
96      2     6:=====+=====
98      3     4:=====+=====
100     1     3:=====+=====
102     3     3:=====+=====
104      0     2:=====+=====
106      1     2:=====+=====
108      0     1:=====+=====
110      0     1:=====+=====
112      0     1:=====+=====
114      0     1:=====+=====
116      0     0:=====+=====
118      1     0:=====+=====
>120     7     0:=====+=====
```

one = represents 22 library sequences

inset = represents 1 library sequences

Mirar el histograma de scores esperados y observados

Mirar el E de la secuencia no relacionada con mayor score

Evaluando la estadística (cont)

```
opt      E()
< 20    13      0:-
22      0      0:-
24      1      0:-
26      0      0:-
28      1      3:*
30     10     20:*
32     21     76:- *
34    105    205:---- *
36    272    422:----- *
38    540    697:----- *
40    937    972:----- *
42   1269   1188:----- *
44   1645   1311:----- *
46   1666   1335:----- *
48   1577   1278:----- *
50   1310   1166:----- *
52   1056   1025:----- *
54    851    876:----- *
56    669    732:----- *
58    423    601:----- *
60    419    487:----- *
62    255    390:----- *
64    196    310:----- *
66    181    245:----- *
68    154    193:----- *
70     99    151:----- *
72     74    118:----- *
74     63     92:----- *
76     60     72:----- *
78     47     56:----- *
80     48     43:----- *
82     36     33:----- *
84     33     26:----- *
86     27     20:----- *
88     21     16:----- *
90     18     12:----- *
92     20      9:----- *
94     20      7:----- *
96     17      6:----- *
98      7      4:----- *
100    10      3:----- *
102    11      3:----- *
104    10      2:----- *
106    11      2:----- *
108     7      1:----- *
110    10      1:----- *
112     6      1:----- *
114     4      1:----- *
116    11      0:----- *
118    10      0:----- *
>120   70      0:----- *
-----
```

one = represents 28 library sequences

inset = represents 2 library sequences

Si los histogramas Obs
vs Exp coinciden

Y si el E del mejor
alineamiento no
relacionado es ~1

La estimaciones
estadísticas están
funcionando bien

Buscando homólogos en los límites

- **Secuencias homólogas distantes a menudo no tienen similitud estadísticamente significativa**
- **Secuencias con regiones de baja complejidad pueden tener similitud estadísticamente significativas, aunque no sean homólogas**
- **Secuencias homólogas generalmente son similares sobre toda la longitud de la secuencia o de un dominio**
- **Secuencias homólogas comparten un ancestro común**
 - **Si hay homología entre A y B; entre B y C; y entre C y D, A y D deben ser homólogos, aun cuando no muestren similitud estadísticamente significativa**

Low complexity sequences

- **Secuencias (o sub-secuencias) con bajo contenido de información**
 - AAAAAAAAAAAAAAAAAAAAAA
 - CAACAACAACAACAACAA
- **Secuencias con sesgo en la composición de bases (nucleótidos) o residuos (aminoácidos)**
- **Por el bajo contenido de información y el sesgo en la composición, suelen dar falsos positivos en las búsquedas por similitud**
 - PolyQ: proteínas con trectos de polyglutamina no están relacionadas por ancestría
 - Ej OTX2 (Transcription factor); CREB-binding protein (connects proteins with different functions); MED15/GAL11 (Subunit of the RNA polymerase II mediator complex)

Low complexity sequences

ORIGIN

```
1 maenlldgpp npraklssp gfsandstdf gslfdlendl pdelipngge lgllnsgnlv
61 pdaaskhkhql sellrgsgs sinpgignvs asspvqqglg gqaagqpnsa nmaslsamgk
121 splsqgdssa pslpkqaast sgtpaasqa lnpqaqkvg latsspatsq tpggicmnan
181 fnqthpglln snsghslinq asqggaqvmn aslqaaqrar gaampvptpa mqqasssyla
```

ORIGIN

```
1 mmsylkqppy avnglsltts gmdllhpsvg ypatprkqrr erttfttraql dvlealfakt
61 rypdifmree valkinlpes rvqvwfknrr akcrqqqqqq qnggqnkvrp akkksspare
121 vssesgtsgq ftpsstsyp tiasssapvs iwspasispl sdplstsssc mqrssypmtyt
181 qasgysqgya gtsyfggmd cgsyltpmhh qlpgpgatls pmgtnavtsh lmqspaslst
241 qygasslgf nsttdcldyk dqtaswklfn nadcldykdq tsswkfqvl
```

//

```
721 mnsfnpmisg nvqlpqapmg praaspmnhs vqmnsmsgsvp gmaisprrmp qppnmgaht
781 nmmmaqapaq sqflpqnqfp sssgamsvgm gqppaqtgvs qgqvpgaalp nplnmlgpqa
841 sqlpcppvtq splhptpppa staagmpslq httpgmtpp qpaaptqpst pvsssgqtpt
901 ptpgsvpsat qtqstptvqa aaqaqvtpqp qtpvqppsua tpqssqqqpt pvhaqqpgtp
961 lsqaaasidn rvptpsvas aetnsqqqgp dvpvlemkte tqaedtepd geskgeprse
1021 mmeedlqgas qvkeetdiae qksepmevde kkpevkvevk eeeessngt asqstpsqp
1081 rkkikfpeel rgalmptlea lyrqdpeslp frqpvdqll gipdyfdvnt npmdlstikr
1141 kldtgqyqep wqyvddvwlw fnnawlynrk tsrvykfcsk laevfeqid pvmqslgycc
1201 grkyefspqt lccygkqlct iprdaayysy qnryhfcekc fteiqgenvt lgddpsqpqt
1261 tiskdqfekk kndtldpepf vdckecgrkm hqicvlhydi iwpsgfvcdn clkktgrprk
1321 enkfsakrlq ttrlgnhled rvnkflrrqn hpeagevfvr vvassdktve vkgpmksrfv
1381 dsqemsesfp yrtkalfafe eidgvdvcff gmhvqeygsd cpppntrrvy isylsiahff
1441 rprclrtavy heiligyley vkklgyvtgh iwacppsegd dyifhchppd qkipkprlq
1501 ewykkmlcka faerihdyk difkqatedr ltsakelpyf egdfwpnvle esikeleqee
1561 eerkkeesta asettegsqg dsknakkkn kktknkssi srankkkpsm pnvsnldsqk
1621 lyatmekhke vffvihlhag pvintlppiv dpdpllsdcl mdgrdafllt ardkhwefss
1681 lrrskwstlc mlvelhtqgq drfvytcnec khhvetrwhc tvcedydlci ncyntkshah
1741 kmvkwglgld degssqgepq skspqesrrl siqrqiqlsv hacqcrnanc slpscqkmkr
1801 vvqhtkgckr ktnggcpcvck qlialccyha khcqnckpv pfclnikhkl rqqqihrlq
1861 qaqlmrrrma tmntrnvppq slpsptsapp gtptqqpstp qtpqppaqp pspvsmmpag
1921 fpsvartqpp ttvstgkpts qvpappppaq pppaaveaar qiereaqqq hlyrvninns
1981 mppgrtgmgp pgsqmapvsl nvprpnqvsg pvmpsmppgq wqqaplpqq pmpglprpvi
2041 smqaaavag prmpsvqppr sispsalqdl lrtlkspssp qqqqqvlnil ksnplmaaf
2101 ikqrtakyva nqpgmqppq lqsqpgmqpp pgmhqqpslq nlnamqagvp rpgvppqqqa
2161 mgglnpqgga lnimnpgnnp nmasmnpqyr emlrqlqlq qqqqqqqqqq qqqqqqgsag
2221 maggmaghgq fqqpqqpggy ppamqqqqgm qqlhplqgss mgqmaaamgq lgqmqgpglg
2281 adstoniqa laqrilaaaa mkaqiaspaa ppomsaahm lsaaqaashl paaqiatls
```

```
lmd intlnggssd tadkirihak nfeaalfaks
vta aaannnikpv eqhhinnlkn sgnsannmnv
qqq qqqqqqqqrr qltpqqqqlv nqmkvapipk
ltp qdmeaakevy kihqqllfka rlqqqqaq
mqp pnssannnpl qqqssqntvp nvlqninqif
mte pvkqsfirky inqalrkiq alrdvknenn
nnn dtiatsatpn aaafsqqqna ssklyqmqqq
qaq aqaqaaqaaq aqaqaaqaaq aqaqaaqaaq
akd vevikqlsld asktnlrld vtlnlsneek
tkn enflkevflq rifvkeilek caegifvkl
lrq qqmmannnngn pgttstgnnn niatqqnmqq
qqq qqqqqqqhiyp sstpgvanys amanapgnni
aat pslnktingk vngtrksnti pvtsipstnk
nps plktqtktngt pnpnmktvq spmgaqpsyn
rfk hrqEIFkdsp mdlfmstlgd clgikdeeml
ard qdsidisikd nklvmkskfn ksnrsysial
tss nmdvgnprkr kasvleispq dsiasvlspd
sek qevtneapfl tsgtsseqfn vwdwnnwtsa
```

BLAST compositional adjustment of scoring matrices

Las matrices de scoring (ej BLOSUM62) son derivadas a partir de alineamientos de proteínas globulares, con una **composición de aminoácidos definida**

Pero muchas veces las secuencias “query” tienen composiciones de aminoácidos con sesgos muy marcados y diferentes a las de las secuencias utilizadas para derivar las matrices

Ej proteínas hidrofóbicas, Cys-rich, proteínas codificadas por genomas con alto sesgo (AT rich, GC rich), que afectan los codones más utilizados

En esos casos hay maneras de ajustar las matrices (ej recalcular BLOSUM62 on-the-fly) para que funcionen mejor frente a estos casos

BLAST compositional adjustment of scoring matrices

Query = human insulin NP_000198
Program = blastp
Database = *C. elegans* RefSeq


Option = NO compositional adjustment

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment



```
>[ref|NP_501926.1] UG INSulin related family member (ins-1) [Caenorhabditis elegans]
Length=109

Score = 34.7 bits (78), Expect = 0.009
Identities = 30/100 (30%), Positives = 41/100 (41%), Gaps = 14/100 (14%)

Query 11  LALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGG 70
          LA+L L P P+ A + LCGS L L VC + +R A+
Sbjct 17  LAILLSSPTPSDASIR--LCGSRLLTTLLAVCRNQLCTGLTAFKRSADQSY----- 66

Query 71  GPGAGSLQPLALEGSLQKRG-IVEQCCTSLCSLYQLENYC 109
          A + + L QKRG I +CC CS L+ +C
Sbjct 67  ---APTTRDLFHIHQKRGGIATECCEKCSFAYLKTFC 103
```

Option = conditional compositional score matrix adjustment

```
>[ref|NP_501926.1] UG INSulin related family member (ins-1) [Caenorhabditis elegans]
Length=109

Score = 33.5 bits (75), Expect = 0.020, Method: Compositional matrix adjust.
Identities = 27/100 (27%), Positives = 39/100 (39%), Gaps = 12/100 (12%)

Query 10  LLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELG 69
          LA+L L P P+ A + LCGS L L VC + +R A+
Sbjct 16  FLAILLSSPTPSDASIR--LCGSRLLTTLLAVCRNQLCTGLTAFKRSADQ-----S 65

Query 70  GPGAGSLQPLALEGSLQKRGIVEQCCTSLCSLYQLENYC 109
          P L + ++ GI +CC CS L+ +C
Sbjct 66  YAPTTRDL--FHIHQKRGGIATECCEKCSFAYLKTFC 103
```

BLAST compositional adjustment of scoring matrices

Fig. 1 Examples of alignment extensions yielded by compositional adjustment of the scoring system. The sequences ...

(a)

```
78 ISGAILFEETLFQKNEAGVPMVNLLHNENIIPGIKVDKGLVNIPCTDEE--KSTQGLDGLAERCKEYYKAGA 147
I GAILFE+T+ K ++ L + ++P +K+DKGL ++ + K L L +R E + G
67 ILGAILFEQTMSKIDGKYTADFLWEEKKVLFPFLKIDKGLNDLDADGVQTMKPNPTLADLLKRANERHIFG- 137

148 RFAKWRTVLVIDTAKGKPTDLS-IHETAWGLARYASICQQNRLVPIVEPEI 197
K R+V+ K+ P ++ + E + +A A + L+PI+EPE+
138 --TKMRSVI----KKASPAGIARVVEQQFEVA--AQVVAAG-LIPIIEPEV 179
```

(b)

```
78 ISGAILFEETLFQKNEAGVPMVNLLHNENIIPGIKVDKGLVNIPCTDEE--KSTQGLDGLAERCKEYYKAGA 147
I GAILFE+T+ K + L + + ++P +K+DKGL ++ + + K L L +R+ E + G
67 ILGAILFEQTMSKIDGKYTADFLWEEKKVLFPFLKIDKGLNDLDADGVQTMKPNPTLADLLKRANERHIFG- 137

148 RFAKWRTVLVIDTAKGKPTDLS-IHETAWGLARYASICQQNRLVPIVEPEILADGPHSIEVCAVVTQKVLSC 218
K+R+V+ K+ P ++ + E + +A A ++ L+PI+EPE+ ++ ++ C + + +
138 --TKMRSVI----KKASPAGIARVVEQQFEVA--AQVVAAG-LIPIIEPEVDINNVDKVQ-CEEILRDEIRK 199

219 VFKALQE-NGVLLEGALLKPNMVTAGYECTAKTTTQDVGFLTVRTLRRTVPPALPGVVFLSGGQSEEEAS 287
+ AL E ++V+L+ L P + E T P + VV LSGG S E+A+
200 HLNALPETSNNMLKLT--PTVENLYEEFTKH-----PRVVRVVALSGGYSREKAN 248
```

(c)

```
78 ISGAILFEETLFQKNEAGVPMVNLLHNENIIPGIKVDKGLVNIPCTDEE--KSTQGLDGLAERCKEYYKAGA 147
I GAILFE+T+ K + L + + ++P +K+DKGL ++ + + K + L L +R+ E + G
67 ILGAILFEQTMSKIDGKYTADFLWEEKKVLFPFLKIDKGLNDLDADGVQTMKPNPTLADLLKRANERHIFG- 137

148 RFAKWRTVLVIDTAKGKPTDLS-IHETAWGLARYASICQQNRLVPIVEPEILADGPHSIEVCAVVTQKVLSC 218
K+R+V+ K+ P ++ + E + +A A ++ L+PI+EPE+ ++ ++ ++ ++
138 --TKMRSVI----KKASPAGIARVVEQQFEVA--AQVVAAG-LIPIIEPEVDINNVDKVQCEEILRDEIRKH 200

219 VFKALQENGVLLEGALLKPNMVTAGYECTAKTTTQDVGFLTVRTLRRTVPPALPGVVFLSGGQSEEEAS 287
+ + ++V+L+ L P + + E T P + VV LSGG S E+A+
201 LNALPETSNNMLKLT--PTVENLYEEFTKH-----PRVVRVVALSGGYSREKAN 248
```

a) standard BLOSUM-62 substitution matrix

b) composition-adjusted matrix derived from BLOSUM-62 with unconstrained relative entropy

b) composition-adjusted matrix derived from BLOSUM-62 with relative entropy constrained to 0.566 bits

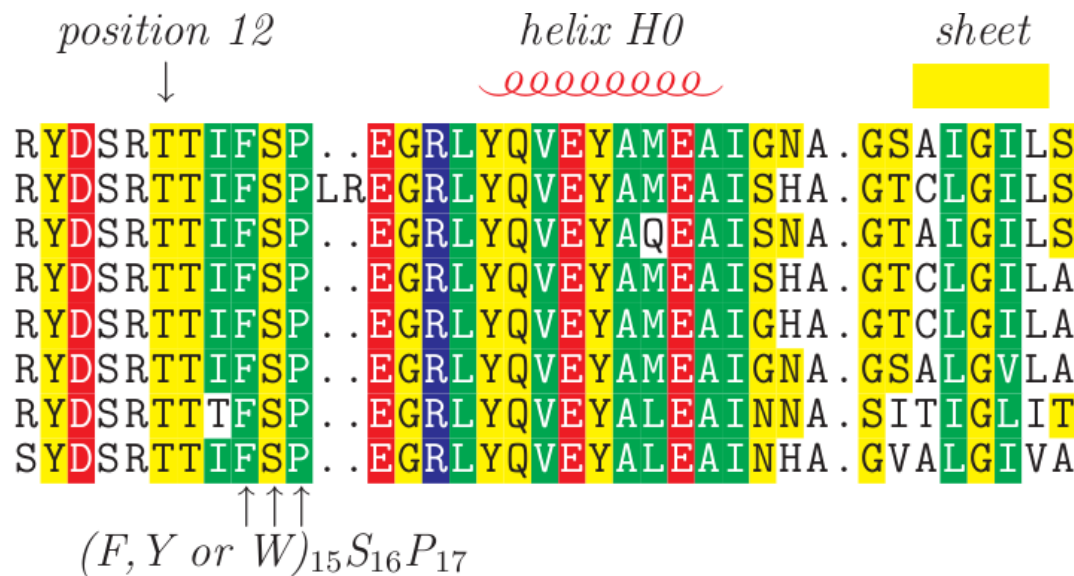
Alineamiento múltiple de secuencias

Algoritmos exactos, heurísticas,
información contenida en un alineamiento

Fernán Agüero

Instituto de Investigaciones Biotecnológicas

Universidad Nacional de San Martín



Alineamientos múltiples

Qué es un alineamiento múltiple?

Example: A multiple sequence alignment corresponding to the WW domain
(Source: SMART database)

O54971/1-33	PLPPGWEKRT	DSN-GRVYFV	N---HNTRIT	QWEDPRS
O43165/1-33	GLPSGWEERK	DAK-GRYYV	N---HNNRTT	TWTRPIM
NED4_HUMAN/1-33	PLPPGWEERT	HTD-GRIFYI	N---HNIKRT	QWEDPRL
O14326/1-33	PLPSGWEML	TNS-ARVYFV	D---HNTKTT	TWDDPRL
O43165_2/1-33	FLPPGWEML	APN-GRPFFI	D---HNTKTT	TWEDPRL
PIN1_HUMAN/1-34	KLPPGWEKRM	SRSSGRVYYF	N---HITNAS	QWERPSG
NED4_HUMAN_1/1-0	PLPPGWEERQ	DIL-GRYYV	N---HESRRT	QWKRPTP
O75853/1-33	PLPPGWEVRS	TVS-GRIFYFV	D---HNNRTT	QFTDPRL
PUB1_SCHPO_2/1-0	RLPPGWERRT	DNL-GRYYV	D---HNTRST	TWIRPNL
YA65_CHICK/1-33	PLPPGWEAK	TPS-QQRYFL	N---HIDQTT	TWQDPK
I83196_2/1-33	GLPPGWEKQ	DDR-GRSYYV	D---HNSKTT	TWSKPTM
YA65_MOUSE/1-33	PLPDGWEQAM	TQD-GEVYYI	N---HKNKTT	SWLDPRL

- Alineamiento de 3 o más secuencias (DNA o proteína)
- Se asume una relación evolutiva entre las secuencias

Alineamientos múltiples

Importancia: muchos métodos en bioinformática usan alineamientos múltiples como *input*.

La calidad de los alineamientos es clave!

Breve intro a evolución de proteínas

Divergencia evolutiva: muchas especies tienen variantes de la misma proteína, todas con esencialmente la *misma* función molecular, pero con secuencias de amino ácidos *diferentes*.

Homología: genes (ADN) y proteínas que evolucionaron a partir de un mismo gen o proteína ancestral se dice que son *homólogos*.

Homólogos, ortólogos, parálogos

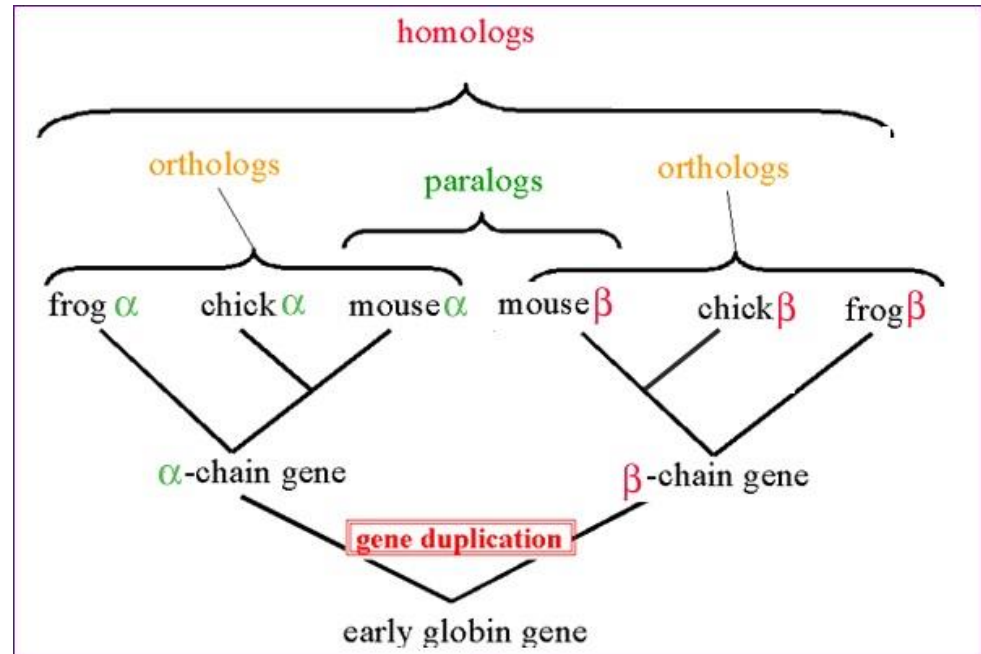
Un par de genes **homólogos** se dice que son **ortólogos** cuando se separaron por especiación.

los genes ortólogos tienden a conservar la misma función molecular en diferentes especies

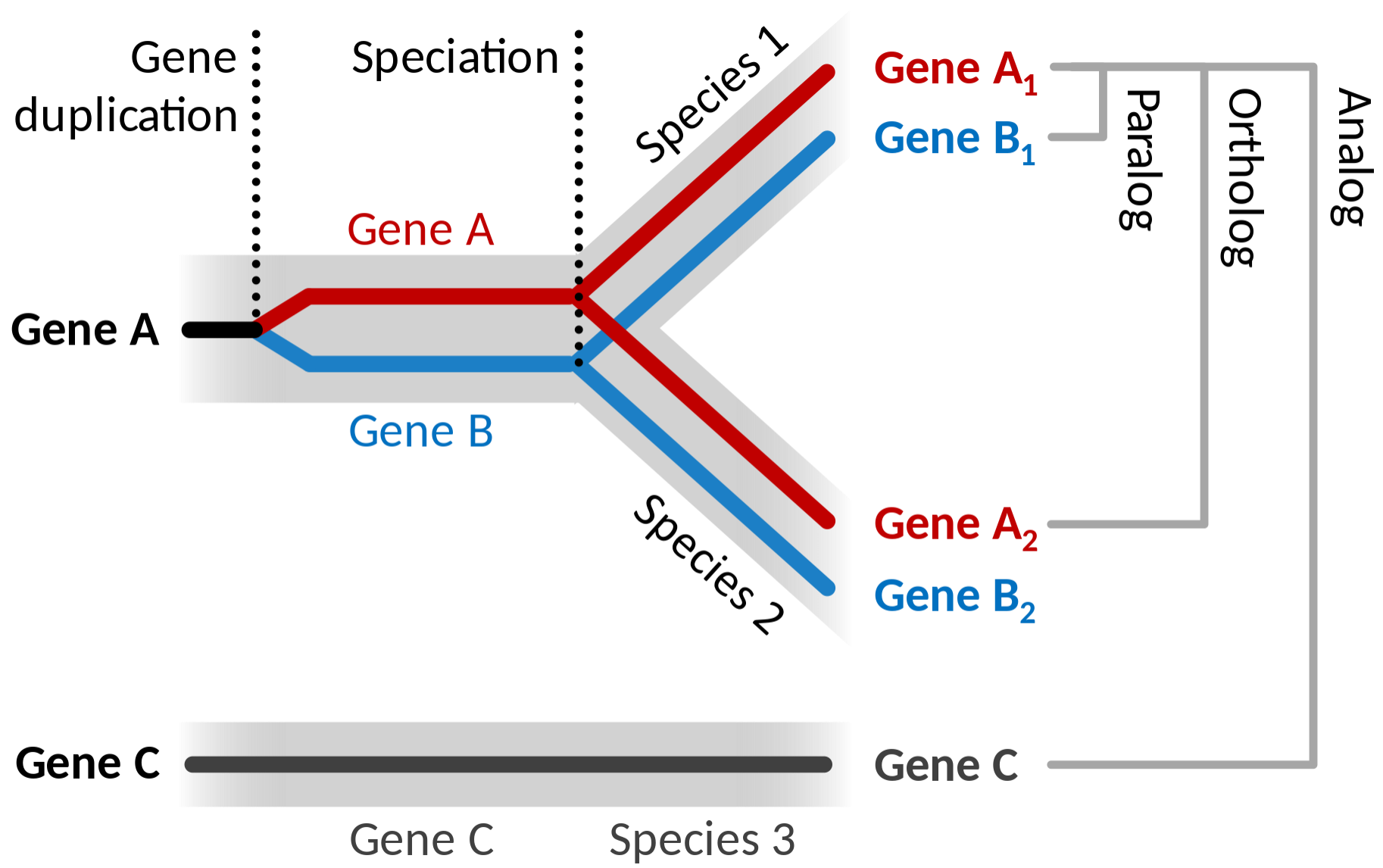
Un par de genes **homólogos** se dice que son **parálogos** cuando se generaron por duplicación dentro de la misma especie.

los genes parálogos tienden a desarrollar o evolucionar hacia funciones diferentes

PERO! Caveat emptor: cuidado con las generalizaciones!



Homólogos, ortólogos, parálogos



By Thomas Shafee - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=70715956>

Box 1: Relationships between genes

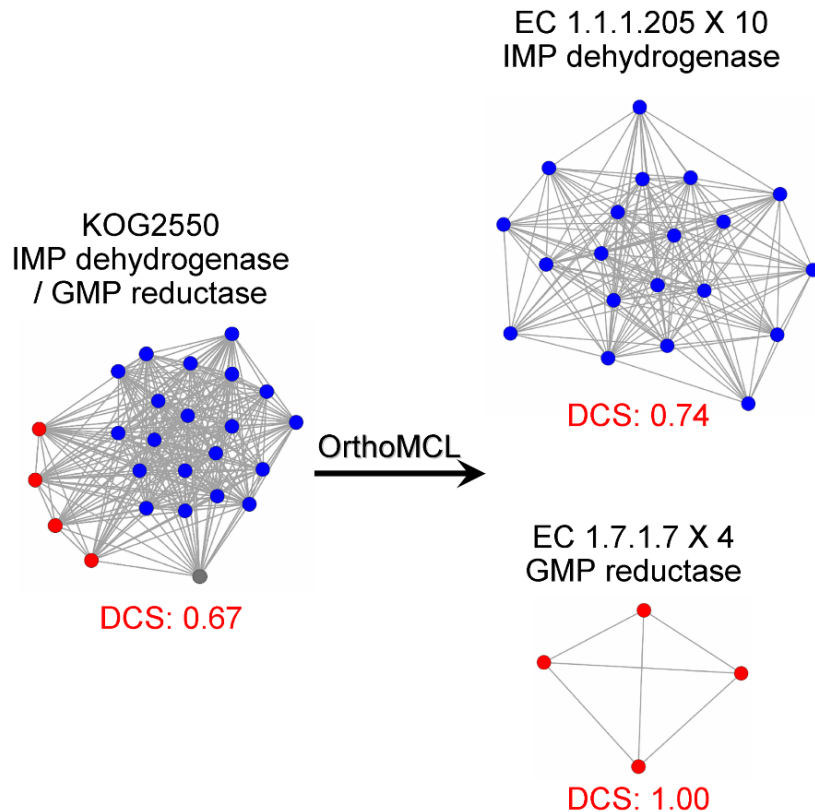
- Homology: genes that share a common origin.
- Analogy: non-homologous genes that perform the same function as a result of convergent evolution.
- Orthology: genes arising by speciation at their most recent point of origin.
- Paralogy: genes arising by duplication at their most recent point of origin.
- Xenology: genes arising by HGT from another organism.
- In-/Out-paralogy: paralogous genes arising from lineage-specific duplication(s) after/before a given speciation event.
- Co-orthology: in-paralogous genes that are collectively, but not individually, orthologous to genes in other lineages (due to their common origin by speciation).
- Orthologous group: collection of all descendants of an ancestral gene that diverged from (after) a given speciation event.

David M. Kristensen, Yuri I. Wolf, Arcady R. Mushegian, Eugene V. Koonin,
Computational methods for Gene Orthology inference,
Briefings in Bioinformatics 12: 379–391.

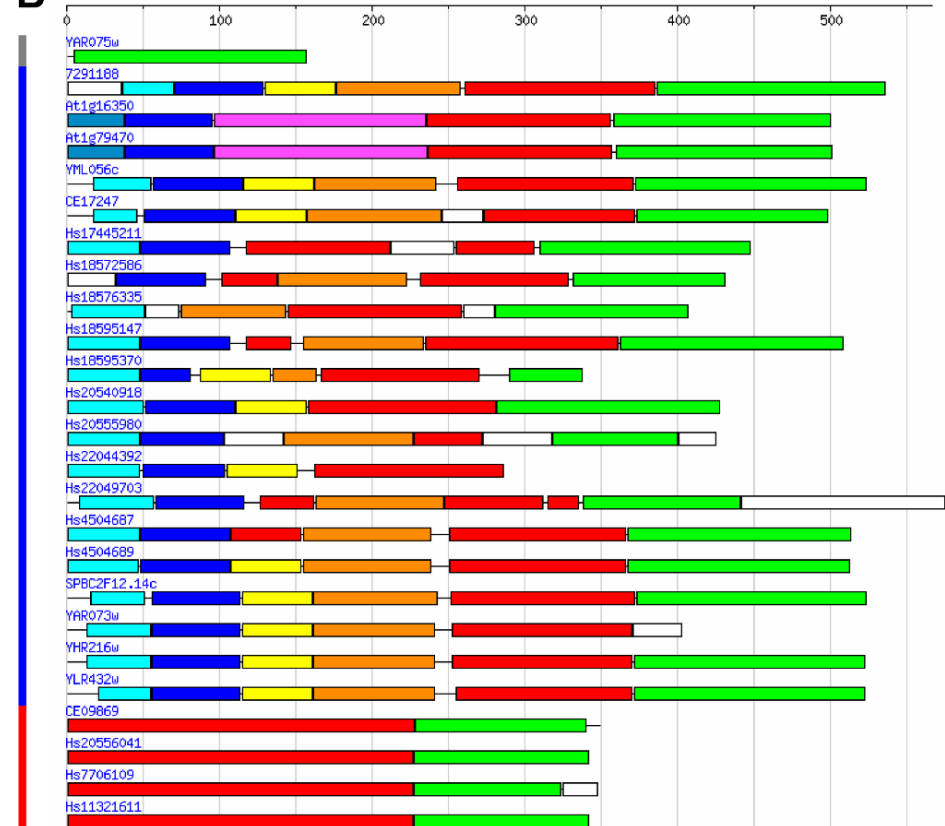
Agrupamiento de Ortólogos

Hay distintas aplicaciones que buscan *agrupar* ortólogos

A



B



Chen F et al 2007 PLOS One 2: e383
DOI: 10.1371/journal.pone.0000383

Y cómo sería un *algoritmo* de alineamiento multiple?

Como en muchos otros algoritmos necesitamos:

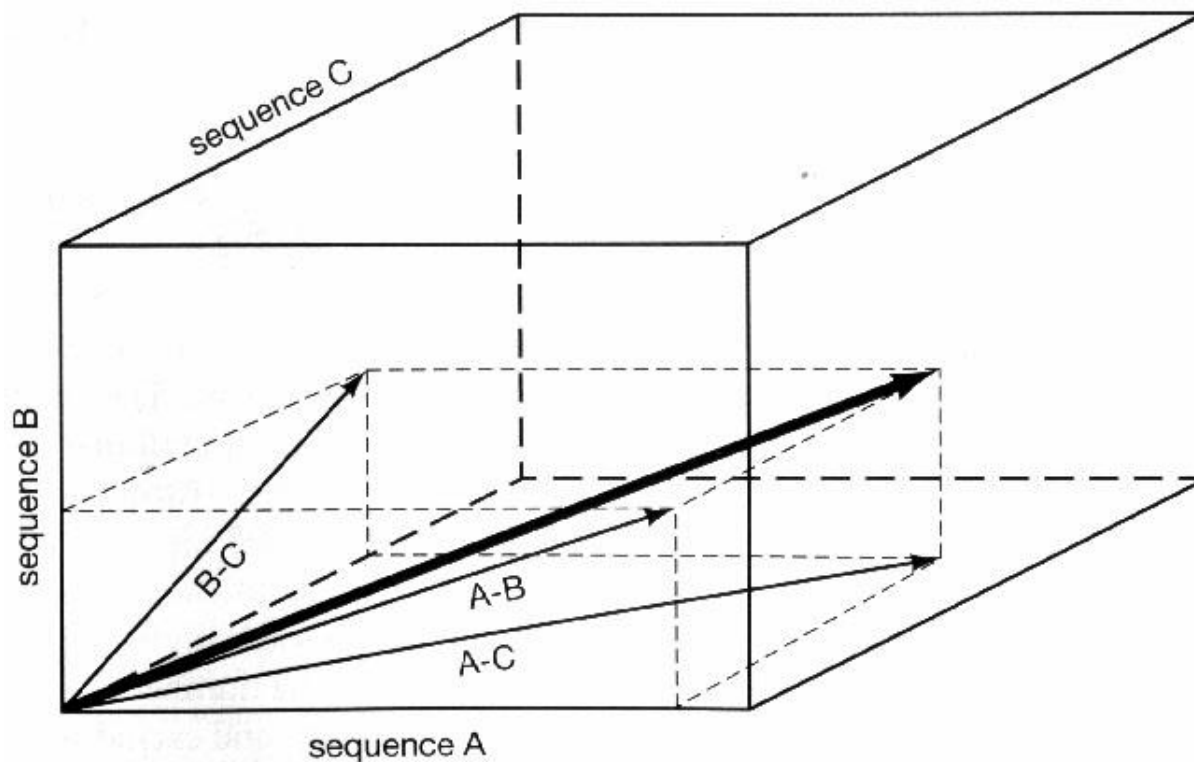
- Una **función objetivo** (métrica)
- Un **procedimiento** para **optimizar** la función objetivo

Para dos alinear **dos** secuencias:

- Una **función objetivo** (métrica)
 - *Sistema de puntajes (Scoring), Matrices (ej BLOSUM62)*
- Un **procedimiento** para **optimizar** la función objetivo
 - *Dynamic programming (ej Needleman-Wunsch)*

Algoritmo exacto

- **Cómo se resuelve un alineamiento múltiple de 3 secuencias?**
- **Usando dynamic programming en una matriz tridimensional**
- **El problema es el mismo: encontrar el camino óptimo en el espacio**



Multiple alignment

```
FHIT_HUMAN  -----MS-F RFGQHLLKP-SVVFL KTELSEALVNRKPVV PGHVLV...
APH1_SCHPO  -----MPKQ LYFSKFVPG-SQVFY RTKLSAAFVNLKPIL PGHVLV...
HNT2_YEAST  MILSKTKKPKSMNKP IYFSKFLVT-EQVFY KSKYTYALVNLKPIV PGHVLI...
Y866_METJA  -----MCIF CKIINGEIP-AKVYV EDEHVLAFLDINPRN KGHTLV...
```

**Un método de alineamiento múltiple verdadero,
alinea todas las secuencias al mismo tiempo.**

**Pero no existe un método computacional que pueda
realizar esto en tiempo razonable para más de 3
secuencias cortas**

Complejidad del algoritmo DP (Dynamic programming)

- **El número de comparaciones que DP tiene que hacer para llenar la matriz (sin usar heurísticas y excluyendo gaps) es el producto de las longitudes de las dos secuencias**
- **La complejidad del algoritmo crece en forma exponencial con el número de secuencias**
- **Alinear dos secuencias de longitud 300 implica realizar 90,000 comparaciones**
- **Alinear tres secuencias de longitud 300 implica realizar 27,000,000 comparaciones**

MSA: global optimal MSAs

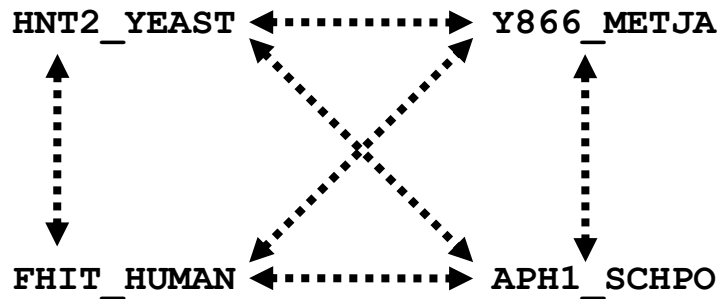
- Needleman-Wunsch o Smith Waterman extendido a una matriz *n-dimensional*
- MSA (Lipman et al. 1989)
 - <http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/msa.html>
 - **Multidimensional dynamic programming**
 - **Usa heurísticas para reducir el espacio de búsqueda**
 - **Varios programas:**
 - msa_50_150 - Alinea no más de 50 secuencias. (c/u < 150 residuos)
 - msa_25_500 - Alinea no más de 25 secuencias (c/u < 500 residuos)
 - msa_10_1000 - Alinea no más de 10 secuencias (c/u < 1000 residuos)
- **Otras heurísticas**
 - **Divide and conquer**
 - Progressive Multiple Sequence Alignments
 - Iterative MSAs ...

MSA: progressive multiple alignments

- Alinear todas las secuencias de a pares
- Usar los scores para construir un árbol filogenético
- Alinear secuencialmente (siguiendo el orden que sugiere el árbol) las secuencias para producir un MSA
- No es un verdadero MSA
- Las secuencias **siempre** se alinean de a pares

MSA: progressive multiple alignments

Align all pairs of sequences.

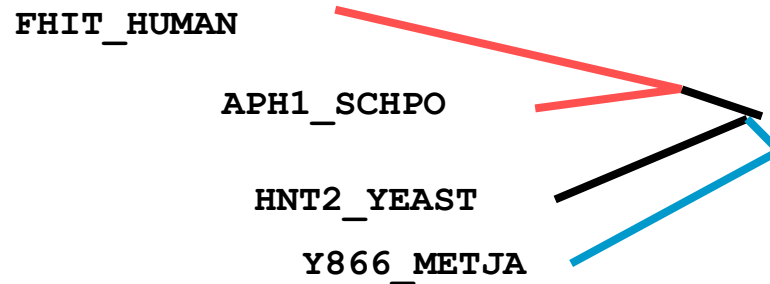


Pairwise alignments: compute distance matrix

	FHIT_HUMAN	APH1_SCHPO	HNT2_YEAST	Y866_METJA
FHIT_HUMAN				
APH1_SCHPO	395			
HNT2_YEAST	316	380		
Y866_METJA	290	300	340	

Progressive multiple alignments

Guide Tree




Pairwise alignments: compute distance matrix

	FHIT_HUMAN	APH1_SCHPO	HNT2_YEAST	Y866_METJA
FHIT_HUMAN				
APH1_SCHPO	395			
HNT2_YEAST	316	380		
Y866_METJA	290	300	340	

Multiple alignment

```
FHIT_HUMAN MSFR FGQHLLKP-SVVFL KTELSEALVNRPVV PGHVLV...
APH1_SCHPO MPKQ LYFSKFPVGSQVFY RTKLSAAFVNLPIL PGHVLV...
HNT2_YEAST MILSKTKKPKSMNKPIYFSKFLVTEQVFYKSKYTYALVNLKPIVPGHVLI...
Y866_METJA MCIF CKIINGEIPAKVVYEDEHVLAFLDINPRNKGHTLV...
```



**Alinear las dos
secuencias más
cercanas**

El alineamiento genera un consenso que se utiliza para alinear las secuencias que quedan.

Desde el punto de vista del alineamiento del primer par, el gap puede insertarse en cualquier lugar

Multiple alignment

```
FHIT_HUMAN  -----MSF RFGQHLLKP-SVVFL KTELSFALVNRKPVV PGHVLV...
APH1_SCHPO  -----MPK QLYFSKFPVGSQVFY RTKLSAAFVNLKPIL PGHVLV...
HNT2_YEAST  MILSKTKKPKSMNK PIYFSKFLVTEQVFY KSKYTYALVNLKPIV PGHVLI...
Y866_METJA  MCIF  CKIINGEIPAKVVYEDEHVLAFDINPRNKGHTLV...
```

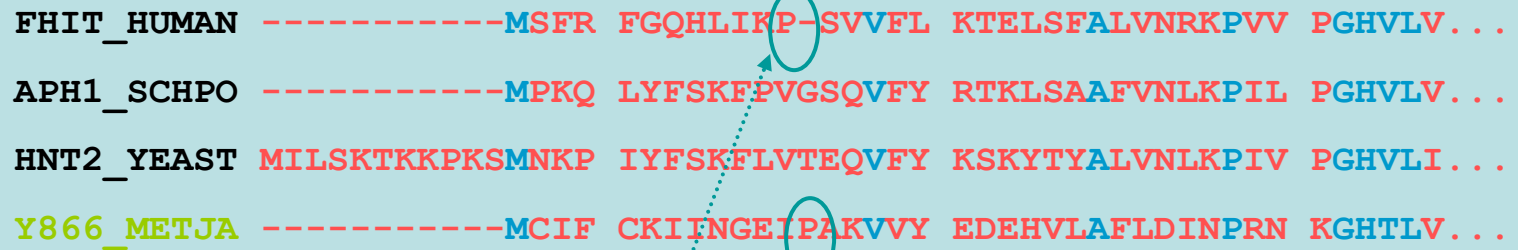


**Alinear las dos
secuencias más
cercanas**

Una vez insertado el gap no
se puede mover porque es
parte del consenso.

Multiple alignment

```
FHIT_HUMAN  -----MSFR FGQHLLKP-SVVFL KTELSFALVNRKPVV PGHVLV...
APH1_SCHPO  -----MPKQ LYFSKEFPVGSQVFY RTKLSAAFVNLKPIL PGHVLV...
HNT2_YEAST  MILSKTKKPKSMNKP IYFSKFLVTEQVFY KSKYTYALVNLKPIV PGHVLI...
Y866_METJA  -----MCIF CKIINGEIPAKVVY EDEHVLAFLDINPRN KGHTLV...
```

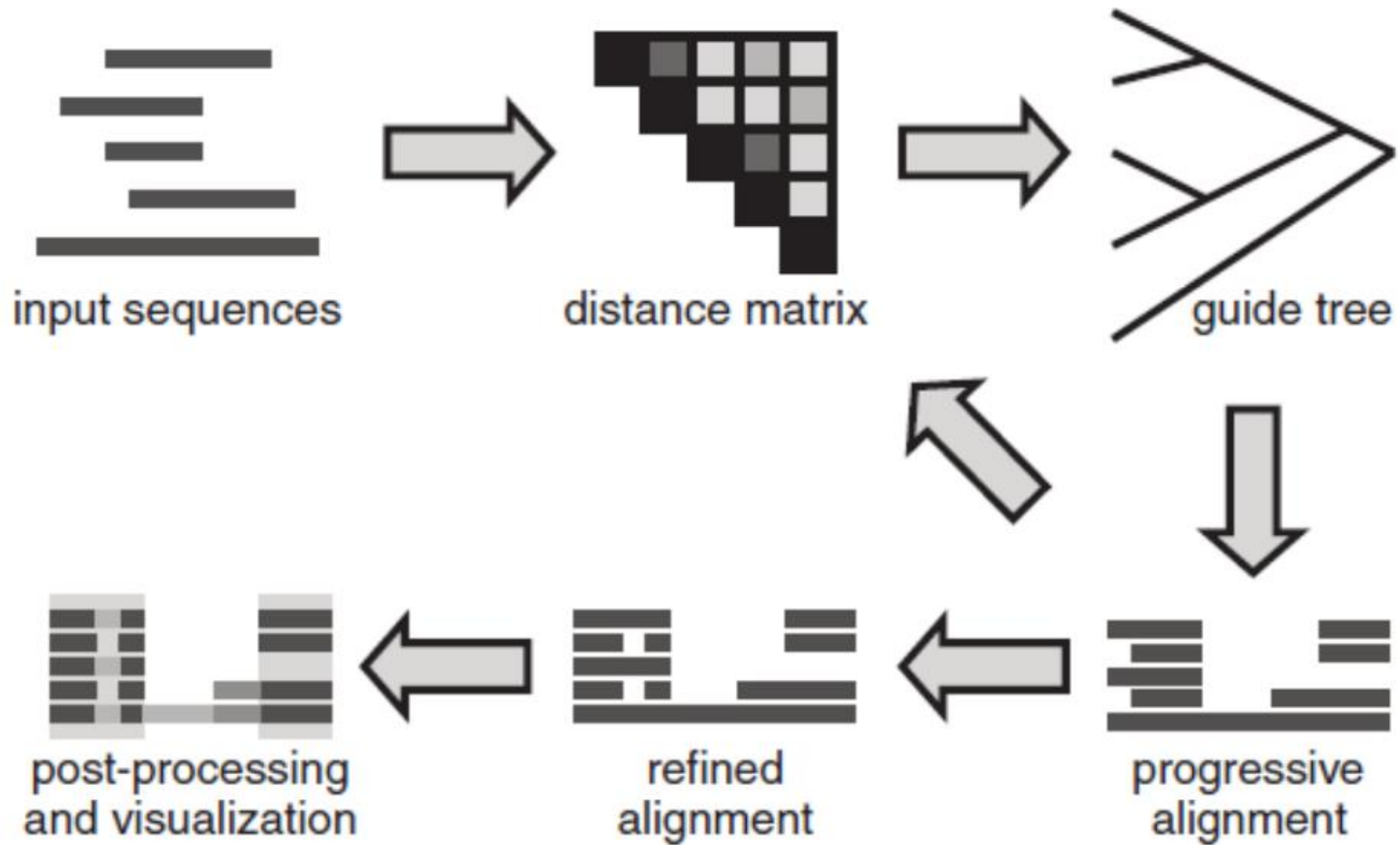


**Alinear la
secuencia
siguiente**

Con suerte, el resultado llegue a ser *similar* al resultado que obtenido por un verdadero método de alineamiento múltiple.

Debido al orden de los alineamientos, la posición del gap no puede cambiarse para alinear estas dos Prolinas (lo cual hubiera resultado en un score mayor.

Resumen de alineamientos progresivos



Clustalw is a progressive multiple alignment tool.

- **Adaptive** gap opening and extension scores
- Choice of DNA or protein gap penalty alignments.
- Available on the web or on PC / Mac / unix.

<http://www.ebi.ac.uk/Tools/msa/clustalw2/> (No longer maintained)

<http://dot.imgen.bcm.tmc.edu:9331/multi-align/options/clustalw.html>

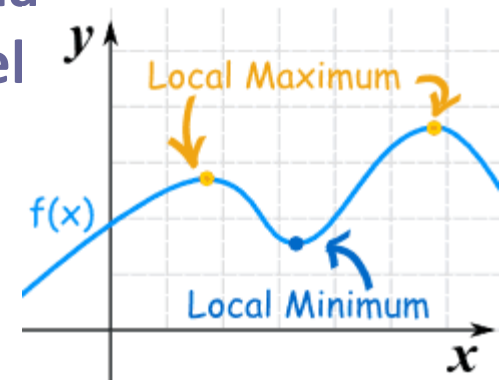
New version ClustalO (Omega)

<http://www.ebi.ac.uk/Tools/msa/clustalo/>

Usa una versión modificada del algoritmo basada en profiles-HMM (se van a ver más adelante en la material)

MSA: métodos iterativos

- Comienzan con un alineamiento multiple inicial
 - Se puede obtener, por ej, usando un método progresivo
- Se optimiza el alineamiento en forma iterativa
- Distintos programas implementan distintas estrategias
- Se realinean subgrupos de secuencias en forma repetida, buscando optimizar el score final del MSA
 - MultAlin (Corpet 1988)
 - PRRP (Gotoh, 1996)
 - DIALIGN (Morgenstern et al. 1996)
 - SAGA (algoritmo genético)
 - MAFFT (Katoh, 2002)
- Como todos los métodos de optimización, pueden quedar atrapados en mínimos locales

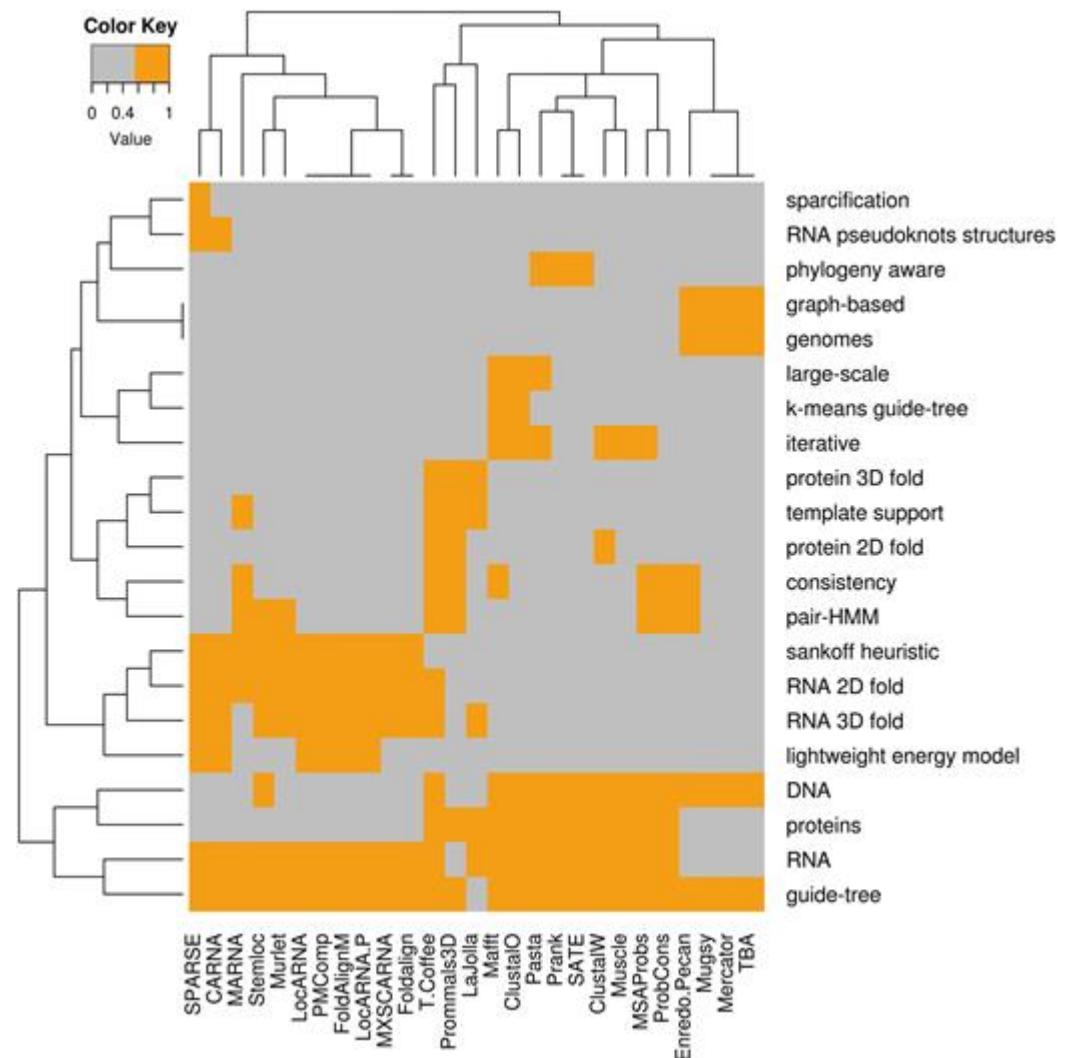


- **SAGA (Notredame & Higgins, 1996)**
 - **Sequence Alignment by Genetic Algorithm**
 - **Genera diferentes MSAs por rearrreglos que simulan inserciones de gaps similares a los que ocurren durante la replicación del DNA**
 - **El proceso continúa hasta que converge en un score que no puede ser mejorado**
 - **Los MSAs no tienen garantía alguna de ser óptimos**
 - **Sin embargo, los alineamientos que produce este método son similares a los que se obtienen por otros métodos**

Otros algoritmos más recientes

- **T-Coffee**
- **MUSCLE**
- **MAFFT**
- **ProbCons**

Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, Cedric Notredame, Multiple sequence alignment modeling: methods and applications, *Briefings in Bioinformatics*, Volume 17, Issue 6, November 2016, Pages 1009–1023, <https://doi.org/10.1093/bib/bbv099>



Visualización y Edición de Alineamientos

Herramienta	URL
Jalview	https://jalview.org
SeaView	http://doua.prabi.fr/software/seaview



protdna.mase

File ▾ Edit ▾ Props ▾ Sites ▾ Species ▾ Footers ▾ Search: Goto: Trees ▾ Help

sel=0 310 Seq:26 Asp Pos:337|178 [AK045539.PE1] 402

AB035322 GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGAGAGAGCTGGGGTTC

AB036423.IBA1 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATAATGCTGGAGAAACTTGGGGTT

AB036423.PE2 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATAATGCTGGAGAAACTTGGGGTT

AB094629.IBA2 GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGAGAGAGCTGGGGTTC

AB128049.AIF1 GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAACGGCGATATTGATATGCTGGAGAAACTTGGGGTTC

AF074959.PE1 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATAATGCTGGAGAAACTTGGGGTT

AF109719.PE7 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATAATGCTGGAGAAACTTGGGGTT

AF129756.PE18 GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTTC

AF299325 -----ATGGAGTTTGATCTGAATGGCAATGGTGATATTGATATTATGTCCCTGAAACGAATGCTGGAGAACTTGGGGTTC

AF299327 -----ATGGAGTTTGATCTGAATGGCAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAACTTGGAGTTC

AF299328 -----ATGGAGTTTGATCTGAATGGCAATGGAGATATCGATATTATGTCCCTGAAGCGAATGCTGGAGAACTTGGGGTT

AF348450.PE1 GCCTTCAAGAAGAAATACATGGAGTTTGACCTGAATGAAGATGGAGGTATCGATATCATGTCCCTGAAGCGAATGATGGAGAACTTGGGGTT

AJ506968.AIF1 ATGTTTAAAAATAAATACATGGAGTTTGATCTCAATGATCAAGGAGACATAGACATAATGGGGTTTAAACGGATGCTTGAAAACTTGGAGTG

AK006184.PE1 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATATTATGTCCCTGAAGCGAATGCTGGAGAACTTGGGGTT

AK006562.PE1 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATATTATGTCCCTGAAGCGAATGCTGGAGAACTTGGGGTT

AK022845.PE1 GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AK028955.PE1 GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGGGTTC

AK045539.PE1 GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGGGTTC

AK091912.PE1 GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AK128526.PE1 GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL136566.PE1 GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL157938.C9ORF58 GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL157938.PE2 GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL157938.PE3 GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL157938.PE7 -----ATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL662801.PE45 GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAACTTGGAGTTC

AL662801.PE46 -----ATGGAGTTTGACCTTAATGGAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAACTTGGAGTTC

AL662847.AIF1 GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAACTTGGAGTTC

AL662847.PE40 -----ATGGAGTTTGACCTTAATGGAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAACTTGGAGTTC

)(<-+_



- **BLOCKS**

- *Blocks are ungapped multiple sequence alignments representing conserved protein regions*
- <http://blocks.fhcrc.org/blocks> (no existe más)
- SeqFire, <http://www.seqfire.org/>
- Gblocks, https://home.cc.umanitoba.ca/~psgendb/doc/Castresana/Gblocks_documentation.html

- **Representan regiones conservadas de un MSA global**
- **No incluyen gaps**
- **Una serie de blocks conservados pueden describir la pertenencia o no a una familia**
- **Pueden buscar usando una secuencia**
- **Pueden usar un MSA para generar blocks**

Información representada en un MSA

- **Un MSA contiene información acerca de las secuencias que lo componen**
- **Si representa a una familia de proteínas:**
 - **regiones conservadas**
 - **residuos conservados**
- **Qué cosas podemos hacer con esta información?**
 - **Muchas**
- **Qué cosas no deberíamos hacer con esta información?**
 - **Generar un consenso**

Consensos

- Un consenso derivado de un MSA contiene para cada posición el residuo más frecuente

OPS2_DROME	MERSHLPETP	FDLAHSGP--	RFQ-AQSSGN	GSV---LDNV	LPDMAHLVNP
OPS2_DROPS	MERSLLPEPP	LAMALLGP--	RFE-AQTGGN	RSV---LDNV	LPDMAPLVNP
OPS2_LIMPO	-----	-MANQLSY--	SSLGWPYQPN	ASV---VDTM	PKEMLYMIHE
OPS2_HEMSA	----MTNATG	PQMAYYGA--	ASMDFGYPEG	VSI---VDFV	RPEIKPYVHQ
OPS2_SCHGR	-----	-MVNTTDFYP	VPAAMAYESS	VGLPLLGNV	PTEHLDLVHP
OPS2_PATYE	----MPFPLN	RTDTALVISP	SEFRIIGIFI	SICCIIGVLG	NLLIIIVFAK
Consenso	MERSMLPETP	?MMA?LGP?P	...		

Problemas!

Usos de los MSAs

- **Para extraer / generar**
 - **Patterns/Motifs**
 - **Profiles**
 - **Fingerprints**
 - **Position Specific Scoring Matrices**
 - **HMMs**
- **Para qué extraer / generar patterns, motifs, etc, etc?**
 - **Para clasificar**
 - **Para alinear secuencias**
 - **Para buscar secuencias similares por métodos más sensibles**

Webster's New Collegiate Dictionary:

mo-tif *n*[F, motive, motif] **1 a:** a usu. recurring salient thematic element in a work of art; *esp:* a dominant idea or central theme

- En secuencias biológicas un **motif** es un patrón recurrente (común) en una serie de secuencias relacionadas
- Los MSAs permiten distinguir regiones de evolución lenta (conservadas) y otras de evolución más rápida en un grupo de secuencias
- Cómo describir/representar las características salientes de un motif?

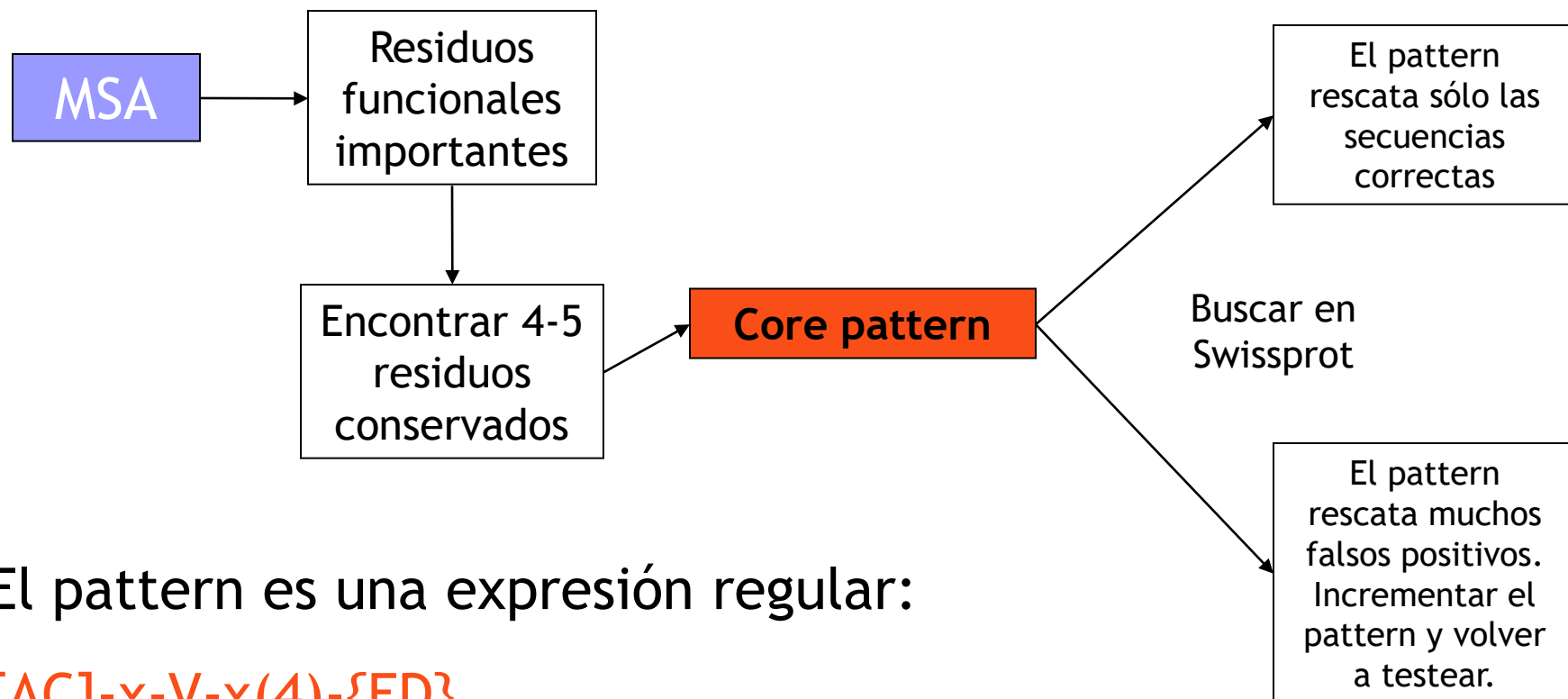
- **Patterns**

- **Descripción (usando una sintaxis particular) de una región corta que tenga relevancia funcional**
- **Cómo se construye un pattern**
 - A partir de la literatura. Se testea contra Swissprot
 - A partir de
 - Enzyme catalytic sites
 - Prosthetic group attachment sites (heme, pyridoxal-phosphate, biotin, etc)
 - Amino acids involved in binding a metal ion
 - Cysteines involved in disulfide bonds
 - Regions involved in binding a molecule (ADP/ATP, GDP/GTP, calcium, DNA, etc.) or another prote



<http://www.expasy.ch/prosite>

Patterns



El pattern es una expresión regular:

[AC]-x-V-x(4)-{ED}

ala/cys-any-val-any-any-any-any-(any except glu or asp)



<http://www.expasy.ch/prosite>

General information about the entry

Entry name [info]	PYRUVATE_KINASE
Accession [info]	PS00110
Entry type [info]	PATTERN
Date [info]	01-APR-1990 CREATED; 01-JUL-1999 DATA UPDATE; 03-AUG-2022 INFO UPDATE.
PROSITE Doc. [info]	PDOC00101

Name and characterization of the entry

Description [info]	Pyruvate kinase active site signature.
Pattern [info]	<code>[LIVAC]-x-[LIVM](2)-[SAPCV]-K-[LIV]-E-[NKRST]-x-[DEQHS]-[GSTA]-[LIVM].</code>

Numerical results [\[info\]](#)

Numerical results for UniProtKB/Swiss-Prot release **2022_03** which contains **568'002** sequence entries.

Total number of hits	84 in 84 different sequences
Number of true positive hits	79 in 79 different sequences
Number of 'unknown' hits	0
Number of false positive hits	5 in 5 different sequences
Number of false negative sequences	22
Number of 'partial' sequences	4
Precision (true positives / (true positives + false positives))	94.05 %
Recall (true positives / (true positives + false negatives))	78.22 %



ScanProsite tool

This form requires to have JavaScript enabled to work correctly.

This form allows you to scan proteins for matches against the [PROSITE collection of motifs](#) as well as against your own patterns.

- ☐ Option 1 - Submit PROTEIN sequences to scan them against the PROSITE collection of motifs.
- ☒ **Option 2 - Submit MOTIFS to scan them against a PROTEIN sequence database.**
- ☐ Option 3 - Submit PROTEIN sequences and MOTIFS to scan them against each other.

[Reset](#)

STEP 1 - Enter a MOTIF or a combination of MOTIFS [Examples](#) [\[help\]](#)

[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV]

Supported input:

- A PROSITE accession e.g. [PS50240](#) or identifier e.g. [TRYPSIN_DOM](#)
- Your own pattern e.g. [P-x\(2\)-G-E-S-G\(2\)-\[AS\]](#)

» [More](#)

» [Options](#) [\[help\]](#)

STEP2 - Select a PROTEIN sequence database [\[help\]](#)

- ☒ UniProtKB
 - ☒ Swiss-Prot ☒ Include isoforms
 - ☐ TrEMBL (sequences belonging to reference proteomes only)
 - ☐ PDB
 - ☐ Your protein database
 - ☐ Randomized UniProtKB/Swiss-Prot
-
- ☐ Exclude fragments (concerns UniProtKB only)

Pattern-Hit Initiated BLAST

Combina búsqueda por motivos (usa sintaxis de Prosite) con BLAST (PSI-BLAST en realidad, ver más adelante)

BLASTP

Enter Query Sequence

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

KIKSRFGHLVKCSMVTNKFGEIPKEAIVGVYMKIHKTEEGEIVGLHQAQFVPEI
QRDCRPFILLSGSELIQVRKEFYDMVDEETRAKIIKMDVDYPSDEDLCSQSF
LKENDYIVFRKDLLRLLEPLNKSFPFIPVQTKKKEIYNHKSFLDLCSEKKVK
QHYPFLAPQKYLPLRVVQAISAPRHKIQELLPPQYKNAGVLJ

Query subrange [?](#)
From
To

Or, upload file No file chosen [?](#)

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

☐ Align two or more sequences [?](#)

Choose Search Set

Databases

☒ Standard databases (nr etc.): New ☐ Experimental databases

Standard

Database ?

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

Exclude Optional ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm

☐ Quick BLASTP (Accelerated protein-protein BLAST)
☐ blastp (protein-protein BLAST)
☐ PSI-BLAST (Position-Specific Iterated BLAST)
☒ PHI-BLAST (Pattern Hit Initiated BLAST)

Enter a PHI pattern [?](#)

☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm [?](#)

 Search database nr using PHI-BLAST (Pattern Hit Initiated BLAST)
☐ Show results in a new window

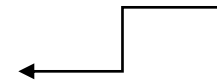
<http://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html>

Profiles

- Representan un MSA en forma de tabla
- Cada posición en el alineamiento corresponde a una fila en el profile
- Para cada posición en el alineamiento el profile contiene la información de frecuencias de aminoácidos que ocurren en esa posición
- Esta información se encuentra representada en forma de scores y penalties e incluye a gaps
- **Un profile no es otra cosa que una serie de matrices de scoring, una para cada posición en el alineamiento**

MSA

1	2	3	4	5	6	7	8	9
---	---	---	---	---	---	---	---	---



1
2
3
4
5
6
7
8

Profile

Un MSA particular

ATP binding RNA helicase ("DEAD" box family)

rhle_ecoli	GVDVLVATPG	RLLDLEHQNAVKLDQV	EILVLDEADR	MLDMGFIHDI
dbp2_schpo	GVEICIAATPG	RLLDMLDSNKTNLRRV	TYLVLDEADR	MLDMGFEPQI
dbp2_yeast	GSEIVIAATPG	RLIDMLEIGKTNLKR	TYLVLDEADR	MLDMGFEPQI
dbpa_ecoli	APHIIVATPG	RLLDHLQKGTVSLDAL	NTLVMDEADR	MLDMGFSDAI
rm62_drome	GCEIVIAATPG	RLIDFLSAGSTNLKRC	TYLVLDEADR	MLDMGFEPQI
p68_human	GVEICIAATPG	RLIDFLECGKTNLRR	TYLVLDEADR	MLDMGFEPQI
rh1b_ecoli	GVDILIGTTG	RLIDYAKQNHINLGAI	QVVVLDEADR	MYDLGFIKDI
yn21_caeel	RPHIIVATPG	RLVDHLENTK	...GFNLKAL	KFLIMDEADR	ILNMDFEVEL
yhm5_yeast	KPHIIVATPG	RLMDHLENTK	...GFSLRKL	KFLVMDEADR	LLDMEFGPVL
me31_drome	KVQLIIATPG	RILDLMDDKVADMSHC	RILVLDEADK	LLSLDFQGML
drsl_yeast	RPDIVIAATPG	RFIDHIRNSA	...SFNVDSV	EILVMDEADR	MLEEGFQDEL
if4a_rabit	APHIIVGTPG	RVFDMNRRYLSPKYI	KMFVLDEADE	MLSRGFKDQI
if41_human	APHIIVGTPG	RVFDMNRRYLSPKYI	KMFVLDEADE	MLSRGFKDQI
vasa_drome	GCHVVIAATPG	RLLDVFVDRTFITFEDT	RFVVLDEADR	MLDMGFSEDM
srm1b_ecoli	NQDIVVATTTG	RLQYIKEENFDCRAV	ETLILDEADR	MLDMGFAQDI
dead_ecoli	GPQIVVGTPG	RLLDHLKRGLDLSKL	SGLVLDEADE	MLRMGFIEDV
if4a_orysa	GVHVVGTPG	RVFDMNRRQLRPDYI	KMFVLDEADE	MLSRGFKDQI
dead_klepn	GPQIVVGTPG	RLLDHLKRGLDLSKL	SGLVLDEADE	MLRMGFIEDV
pl10_mouse	GCHLLVATPG	RLVDMMERGKIGLDFC	KYLVLDEADR	MLDMGFEPQI
p54_human	TVHVVIATPG	RILDLIKKGAKVDHV	QMIVLDEADK	LLSQDFVQIM
if4a_drome	GCHVVVGTPG	RVYDMINRKLRTQYI	KLFVLDEADE	MLSRGFKDQI
ded1_yeast	GCDLLVATPG	RLNDLLERGKISLANV	KYLVLDEADR	MLDMGFEPQI
ms16_yeast	RPNIVIAATPG	RLIDVLEKYS	...NKFFRFV	DYKVLDEADR	LLEIGFRDDL
pr28_yeast	GCDILVATPG	RLIDSLENHLLVMKQV	ETLVLDEADK	MYDLGFEDQV
if4n_human	GQHVVGATPG	RVFDMIRRRSLRTRAI	KMLVLDEADE	MLNKGFKDQI
an3_xenla	GCHLLVATPG	RLVDMMERGKIGLDFC	KYLVLDEADR	MLDMGFEPQI
dbp1_yeast	GCDLLVATPG	RLNDLLERGKVSLANI	KYLVLDEADR	MLDMGFEPQI
if4a_yeast	DAQIVVGTPG	RVFDNIQRRRFRTDKI	KMFILDEADE	MLSSGFKEQI
spb4_yeast	RPQILIGTPG	RVLDFLQMPAVKTSAC	SMVVMDEADR	LLDMSFIKDT
if4a_caeel	GIHVVGATPG	RVGDMINRNALDTSRI	KMFVLDEADE	MLSRGFKDQI
pr05_yeast	GTEIVVATPG	RFIDILTND	.GKLLSTKRI	TFVVMDEADR	LFDLGFEPQI
if42_mouse	APHIVVGTPG	RVFDMNRRYLSPKWI	KMFVLDEADE	MLSRGFKDQI
dhh1_yeast	TVHILVGTPG	RVLDLASRKVADLSDC	SLFIMDEADK	MLSRDFKTI
db73_drome	KADIVVTTPG	RLVDHLHATK	...GFCLKSL	KFLVILDEADR	IMDAVFQNW
yk04_yeast	GCNFIIGTPG	RVLDHLQNTK	VIKEQLSQSL	RYIVLDEGDK	LMELGFDETI
ybz2_yeast	SGQIVIAATPG	RFLELLEKDN	.TLIKRFSKV	NTLILDEADR	LLQDGHFDEF
yhw9_yeast	KPHFIIATPG	RLAHHIMSSG	DDTVGGMLRA	KYLVLDEADI	LLTSTFADHL
glh1_caeel	GATIIIVGTVG	RIKHFCEEGTIKLDKC	RFFVLDEADR	MIDAMGFGTD

Un profile generado a partir del MSA

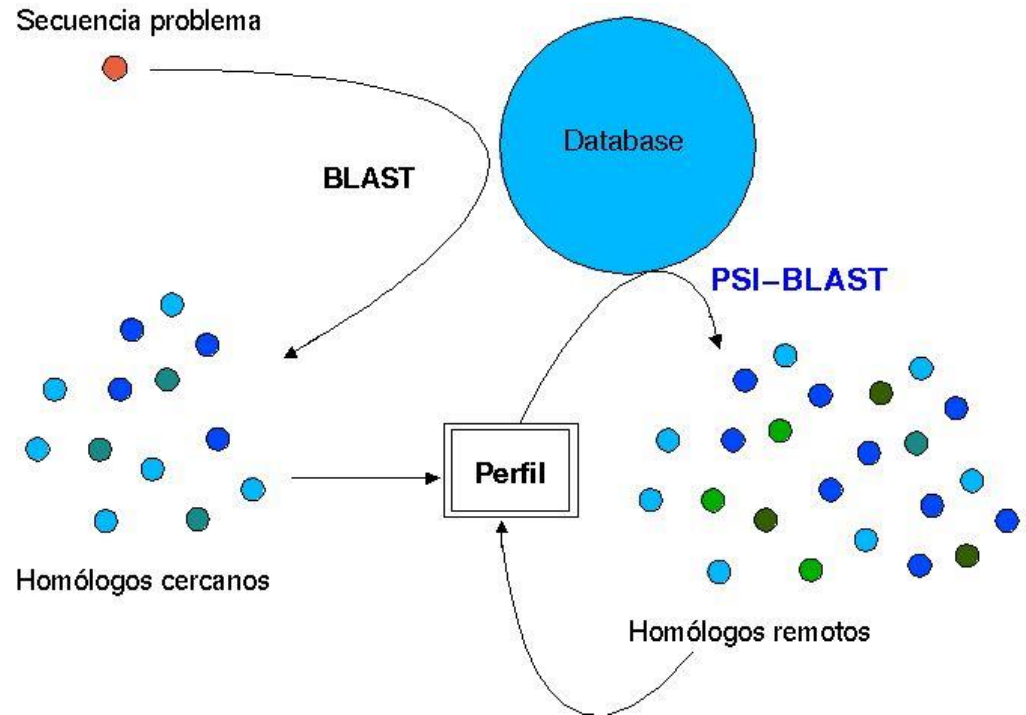
Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	Gap	Len	..
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11	100	100	
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1	100	100	
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27	100	100	
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11	100	100	
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8	100	100	
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9	100	100	
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10	100	100	
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10	100	100	
P	31	6	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48	12	100	100	
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30	100	100	
! 11																									
R	-30	10	-30	0	0	-50	-30	50	-30	80	-40	20	10	30	40	150	10	-10	-30	140	-60	20	100	100	
L	-2	-17	-15	-18	-12	38	-13	-9	38	-12	49	39	-15	-9	-9	-15	-11	0	38	6	12	-10	100	100	
L	0	-12	-15	-14	-9	32	-12	-7	32	-7	41	35	-11	-9	-6	-12	-9	0	29	6	9	-7	100	100	
D	15	58	-27	78	54	-52	35	27	-12	16	-26	-21	38	6	41	3	9	10	-12	-57	-25	50	100	100	
L	-5	-5	-7	-8	-4	24	-12	13	13	-6	25	17	-1	-7	0	-2	-8	-3	10	11	17	-2	100	100	
L	3	-13	-13	-13	-8	31	-11	-8	34	-9	41	36	-12	-7	-5	-13	-8	2	31	-1	8	-6	100	100	
E	6	19	-15	23	27	-21	9	15	-6	18	-8	-1	16	6	23	12	6	5	-6	-15	-16	25	100	100	
K	3	14	-12	11	12	-16	2	10	-5	23	-7	4	15	6	15	22	8	3	-5	7	-15	14	100	100	
G	11	17	0	16	14	-16	19	5	-6	11	-11	-5	16	9	8	4	14	15	-1	-13	-14	11	100	100	
T	12	9	-1	7	7	-8	9	2	4	12	0	4	10	5	4	3	9	12	7	-8	-8	5	100	100	
! 21																									
D	1	1	0	2	1	-1	1	0	1	0	0	0	1	0	1	0	0	1	2	-3	-1	1	22	22	
T	2	2	0	3	2	-2	3	0	2	0	0	0	1	1	1	-1	1	4	2	-5	-2	2	22	22	
K	0	1	-3	0	1	0	0	0	1	4	1	3	1	0	1	1	0	3	1	0	-2	1	22	22	
G	3	3	0	4	4	-1	6	-1	3	0	1	1	3	1	1	-2	4	3	5	-6	-3	2	22	22	
L	5	-6	-4	-7	-4	16	-2	-4	21	-4	23	17	-5	-4	-4	-8	-2	4	19	0	6	-4	22	22	
B	5	16	-6	15	11	-15	10	6	-3	16	-8	-1	15	4	9	10	12	7	-2	-3	-11	10	100	100	
L	1	-13	-12	-14	-9	27	-8	-7	24	-8	36	30	-10	-5	-7	-10	-4	7	23	6	9	-8	100	100	
D	7	19	-7	22	17	-22	13	7	-6	19	-11	-3	14	8	15	14	17	6	-5	-5	-18	16	100	100	
K	11	10	-3	10	9	-12	5	9	-4	16	-6	0	10	6	11	12	10	4	-4	3	-8	10	100	100	
V	7	-10	11	-11	-10	14	0	-8	31	-11	19	16	-10	0	-10	-12	2	8	34	-22	9	-10	100	100	
K	8	9	-4	9	9	-13	11	1	0	16	-4	4	8	7	8	11	13	12	3	-2	-15	8	100	100	
L	3	4	-9	3	6	3	-2	8	9	7	10	10	5	0	8	3	0	5	7	-2	0	7	100	100	
L	1	-13	-13	-13	-9	32	-11	-7	32	-9	42	36	-12	-7	-6	-13	-9	3	33	2	8	-7	100	100	
*	99	0	25	208	120	94	137	44	181	105	256	94	41	62	64	144	59	99	162	3	35	0			

Usos de los profiles

- **También conocidos como**
 - **Position-Specific Scoring Matrix (PSSM)**
- **Derivación de motifs (patterns)**
- **Generación de un MSA**
 - **partiendo de un MSA que se supone representativo de una familia o grupo de proteínas, se genera un profile**
 - **el profile se usa para generar alineamientos nuevos con proteínas no representadas originalmente en el profile**
 - **Más sensible que una matriz de scoring sitio-inespecífica**
- **Búsqueda de secuencias similares en bases de datos**
 - **El 'query' no es una secuencia, sino el profile**

Position-Specific-Iterated BLAST

1. La 1ra iteración es un BLAST tradicional
2. A partir de los hits se calcula un MSA y a partir del MSA se deriva un profile (PSSM)
3. A partir de la segunda iteración, se usa la PSSM como query



<ftp://ftp.ncbi.nih.gov/blast/documents/blastpgp.html>

Profile HMMs

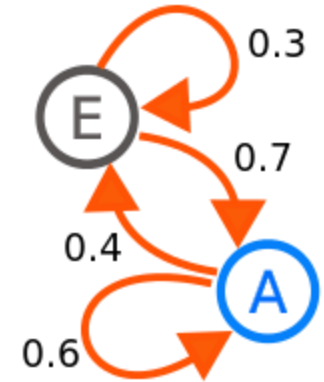
- La información contenida en un profile puede representarse de otras formas
- Los profiles originales contienen scores y penalidades basados en las frecuencias de ocurrencia
- Un profile (o un MSA) también puede representarse como una cadena de eventos con probabilidades de ocurrencia (Markov Chain)
- **Veamos un ejemplo!**

Markov Chains: una pequeña intro

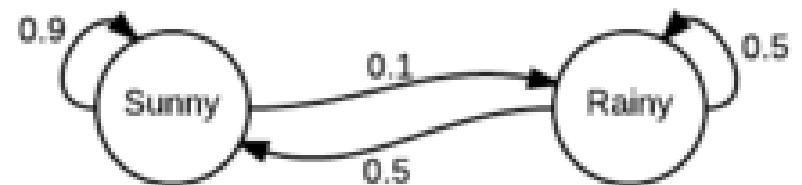
Una cadena de Markov es un sistema matemático que *transita* entre distintos *estados*, de acuerdo a probabilidades

Es un proceso azaroso y sin memoria

El próximo estado del sistema sólo depende del estado actual y no de la secuencia de estados precedentes (historia)

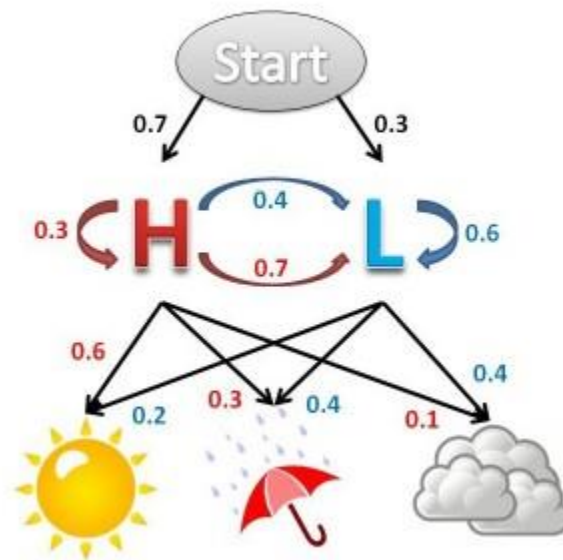
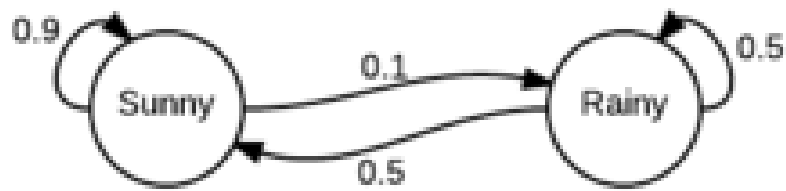


Markov Chain, Wikipedia. http://en.wikipedia.org/wiki/Markov_chain



Hidden Markov Models

Un modelo de Markov es un modelo probabilístico de algún Sistema, en donde existen estados no observables (ocultos).



Profile HMMs

El modelo se
inicia con
transiciones
equiprobables

Y se **entrena** con
un alineamiento

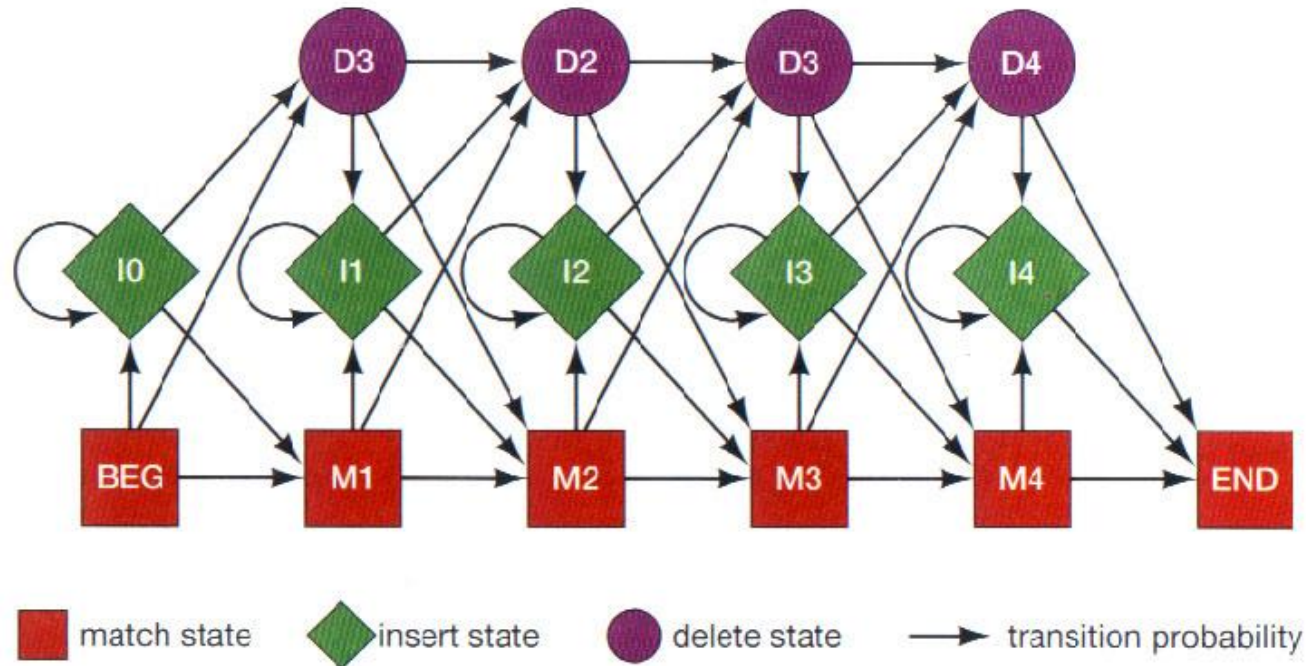
El profile ahora
está codificado en
forma de **estados**
y **probabilidades**
de transición

A. Sequence alignment

N	•	F	L	S
N	•	F	L	S
N	K	Y	L	T
Q	•	W	-	T

RED POSITION REPRESENTS ALIGNMENT IN COLUMN
GREEN POSITION REPRESENTS INSERT IN COLUMN
PURPLE POSITION REPRESENTS DELETE IN COLUMN

B. Hidden Markov model for sequence alignment



- **HMMER**

- <http://hmmer.org>



- **Paquete de programas para trabajar con profile HMMs**

- **genera profile HMMs a partir de MSAs**
 - **usa los HMMs para realizar búsquedas en bases de datos de secuencias**
 - **puede buscar en bases de datos de profile HMMs a partir de una secuencia**



- Una base de datos de profile HMMs
- (y de MSAs)
 - Wellcome Trust Sanger Institute
 - Stockholm Bioinformatics Centre
 - Janelia Farm
- Representan dominios proteicos
- Pueden buscar
 - a partir de palabras clave
 - a partir de una secuencia
- Pfam 35.0 (Noviembre 2021, 19632 families)



Sequence information

Alignment

☒ Seed (12) ☐ Full (28)

Format:

Hyperlinked plain text

[Retrieve alignment](#)

Visualize domain structures

☒ Seed (12) ☐ Full (28)

display per page.

[Retrieve domain structures](#)

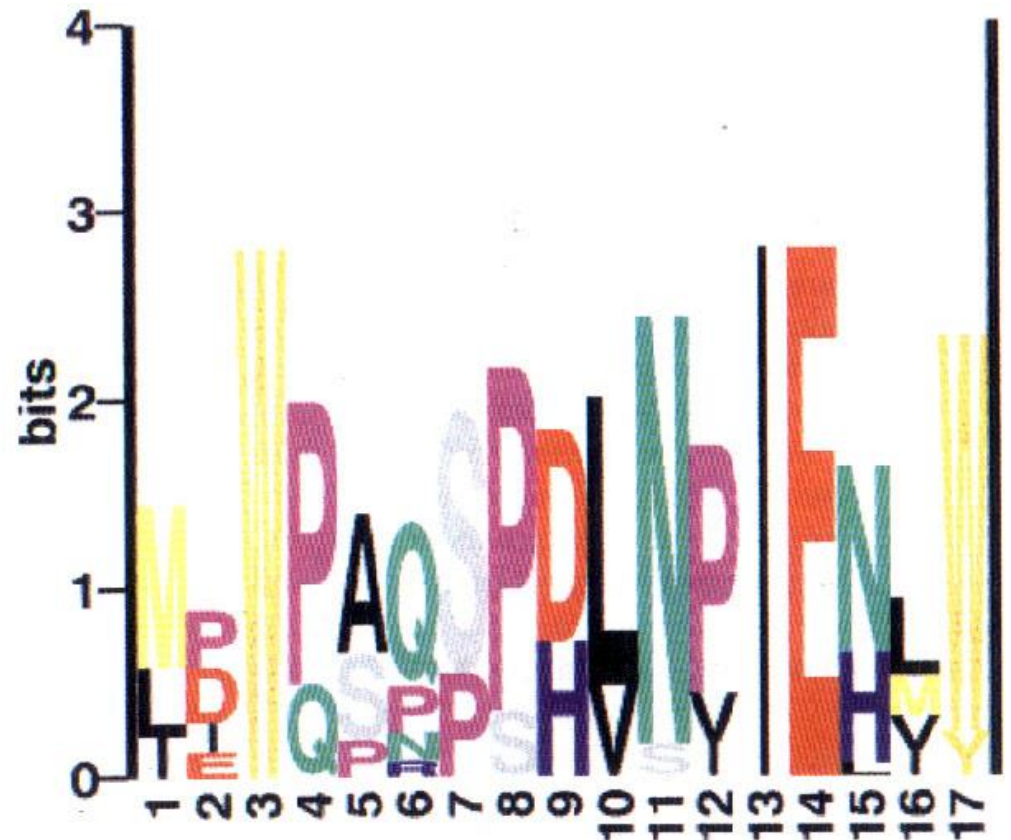
Species distribution

Tree depth:

[View species tree](#)

- Los motifs se pueden representar de distintas maneras (patterns por ejemplo)
- Sin embargo, los patterns no les dan peso a las distintas sustituciones
- [AC]-x-V-x(4)-{ED}
- Una **Position Specific Scoring Matrix** es una descripción de un motif en términos de una matriz

- Evaluar la información que contiene una PSSM usando Sequence Logos
- <http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html>



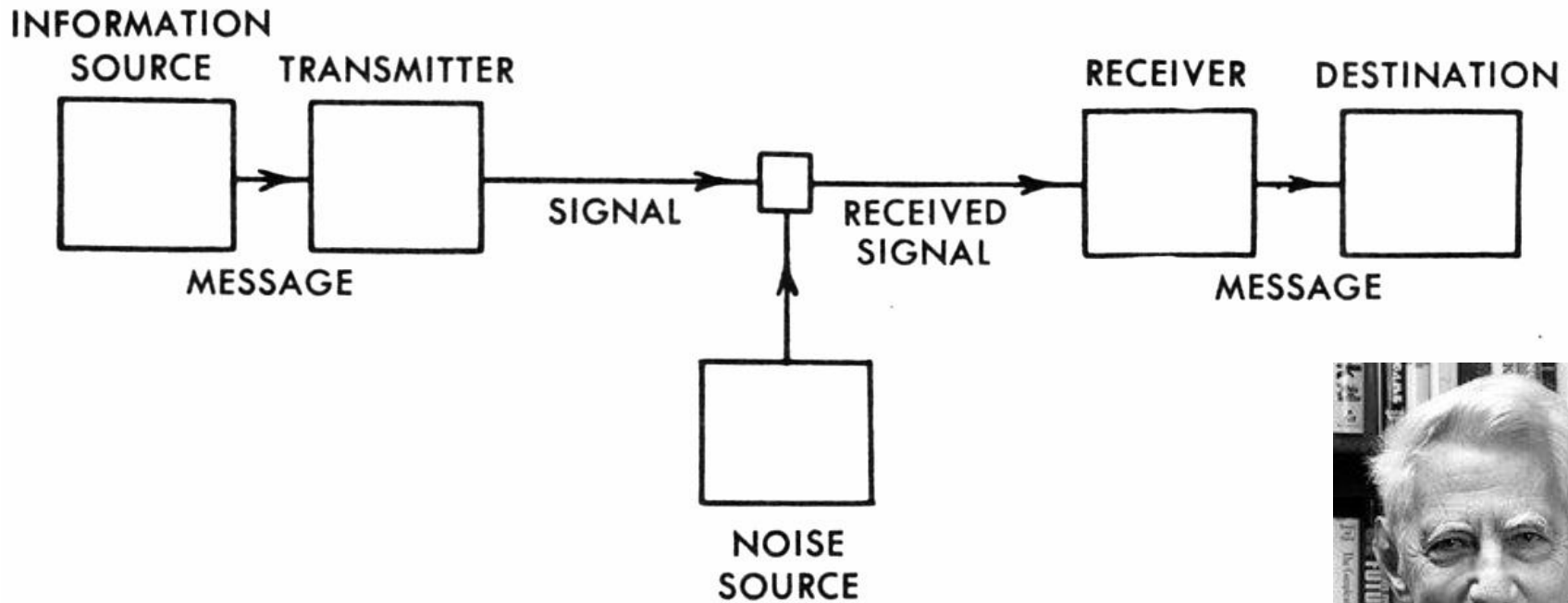
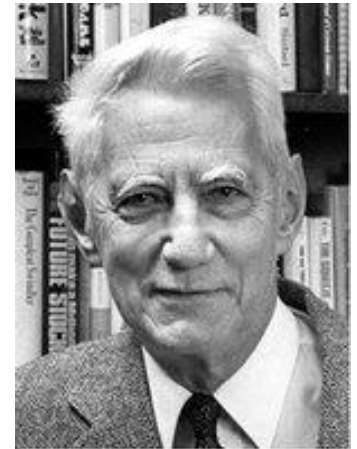


Fig. 1. — Schematic diagram of a general communication system.

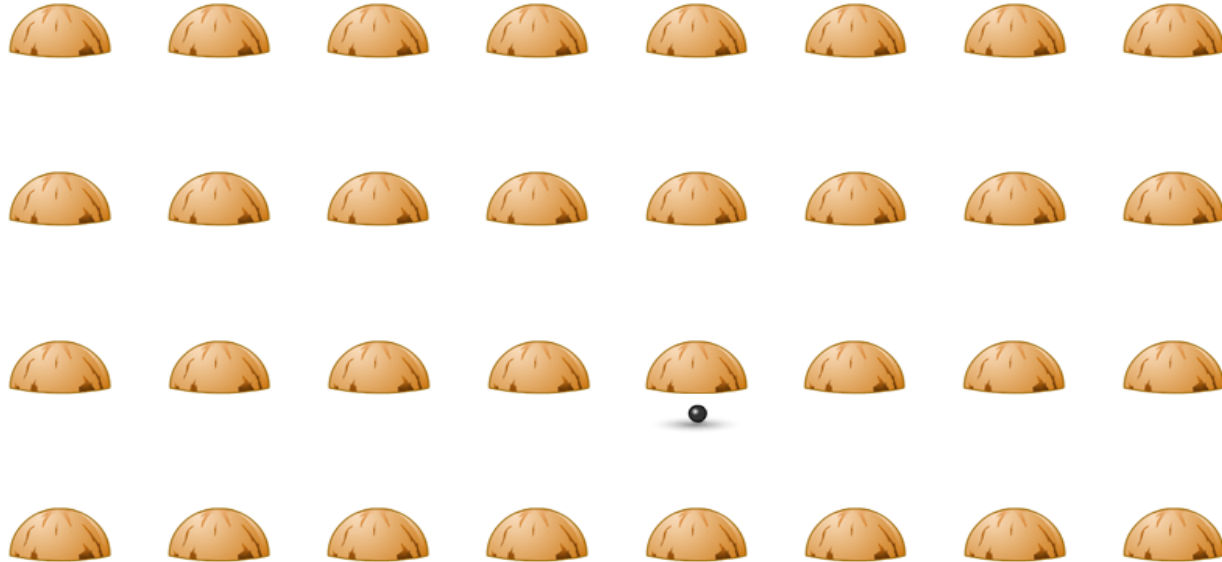


Claude Shannon

Information theory

- **Entropía:** medida de desorden de un sistema
- **La termodinámica** provee herramientas para calcular entropía
- **El desorden implica falta de *información* sobre el estado exacto de un sistema**
- **Claude Shannon / Leon Brillouin**
 - **Information theory**
 - **La Información es una combinación de**
 - Certain + Uncertain, Expected + Unexpected
 - **El grado de *sorpresa* que genera un evento que ya ocurrió es *cero***
 - **Si se reporta un evento poco probable, la información que se provee es *mayor***
 - **La información se incrementa cuando la probabilidad baja**

Shell game



Shell Game (Thimblorig)

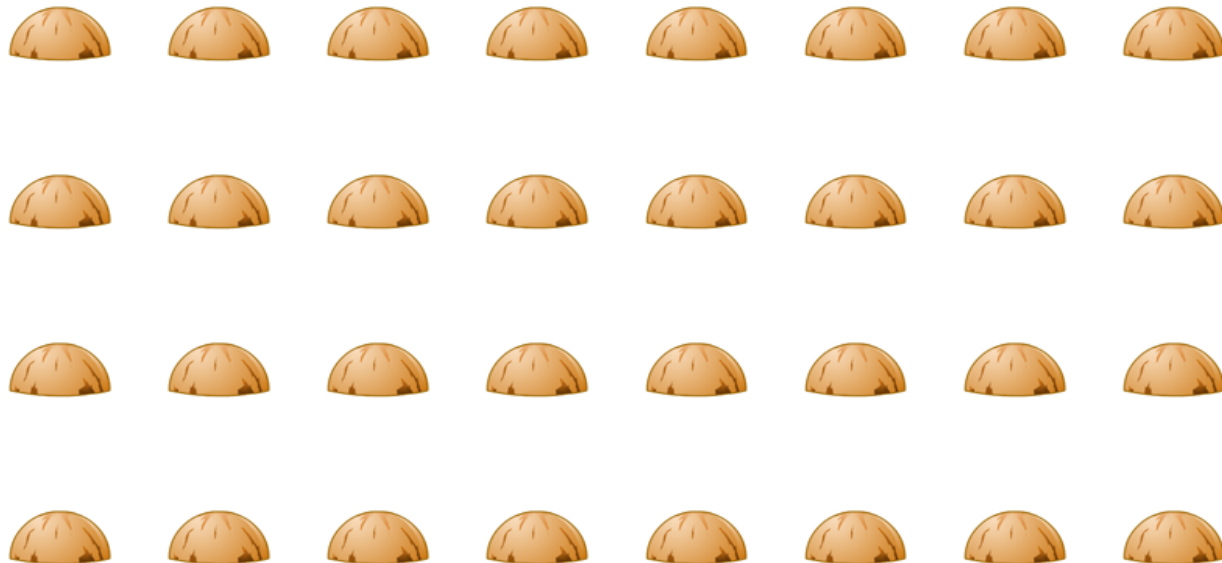
Adivinar en qué taza / nuez
está escondida la bolita.

Uncertainty

Si hay 64 nueces, cuántas
preguntas hay que hacer
para llegar a la respuesta?

Probability

$$p(object) = 1/64$$



Shell game



Shell Game (Thimblergig)

Adivinar en qué taza /
nuez está escondida
la bolita.

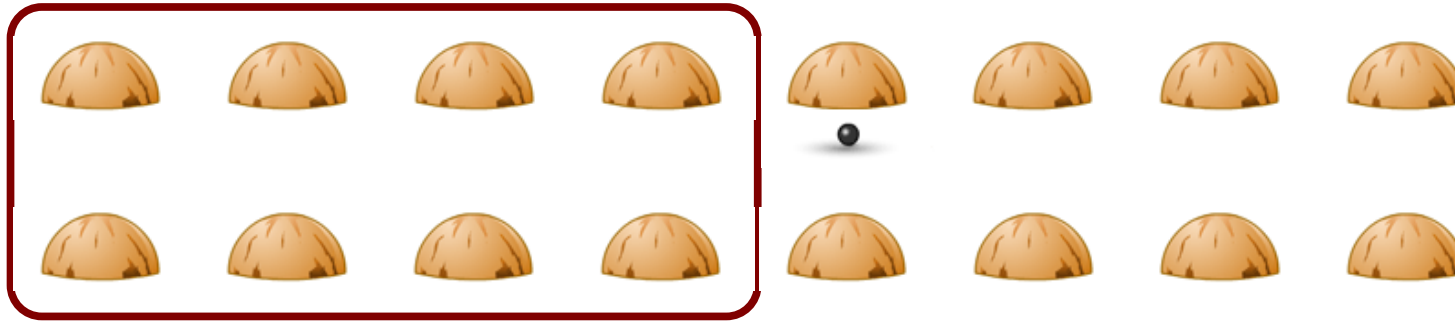
Uncertainty

Si hay 64 nueces,
cuántas preguntas
hay que hacer para
llegar a la respuesta?

Probability

$$p(\text{object}) = 1/64$$

Shell game



Shell Game (Thimblrig)

Adivinar en qué taza /
nuez está escondida
la bolita.

Uncertainty

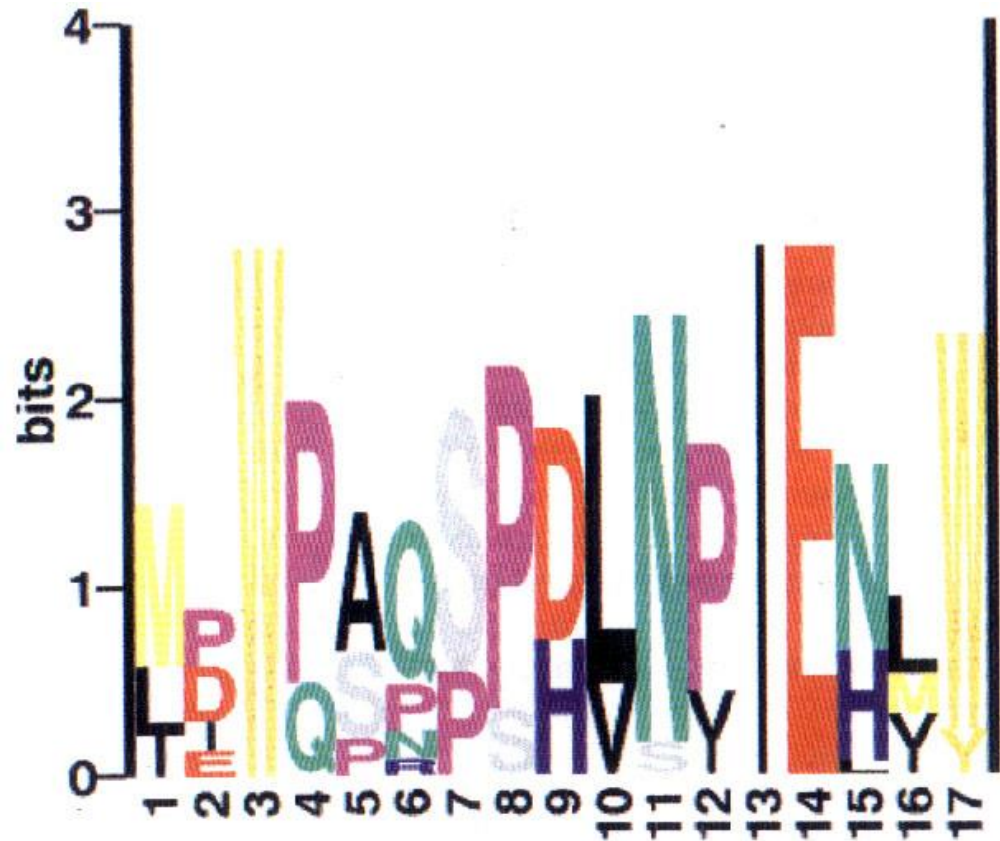
Si hay 64 nueces,
cuántas preguntas
hay que hacer para
llegar a la respuesta?

Probability

$p(object) = 1/64$

- Las preguntas secuenciales reducen las posibilidades (incertidumbre) de 64 a 32, luego a 16, 8, 4, 2, y finalmente 1.
- 6 preguntas son suficientes (peor caso) para encontrar la bolita.
- Esta es una manera de cuantificar la incertidumbre
- La incertidumbre también se puede calcular a partir de las probabilidades
 - Uncertainty = $-\log_2(1/64) = 6$

- **Information content of a PSSM**
 - Objetivo: conocer qué residuo pertenece a cada columna en el motivo
 - 20 residuos (20 posibilidades),
 $\log_2(20) = 4.32$
- **Sequence Logos**
 - Forma de visualización desarrollada por Tom Schneider
 - Grafica la cantidad de información (*disminución en la incertidumbre*) que nos da la matriz para cada posición





- **Protein Fingerprints DB**

- <http://www.bioinf.man.ac.uk/dbbrowser/PRINTS>

- **Qué es un fingerprint?**

- Una serie de motifs conservados en un orden particular
 - Se utilizan para predecir la ocurrencia de motifs similares en una secuencia
 - Importa la presencia y el orden de los motifs
 - Una proteína de la misma familia tiene todos los motifs en orden.
 - En el caso de una superfamilia, miembros de distintas familias pueden tener matchs parciales contra el fingerprint

SUMMARY INFORMATION

9 codes involving 8 elements
 0 codes involving 7 elements
 10 codes involving 6 elements
 29 codes involving 5 elements
 5 codes involving 4 elements
 4 codes involving 3 elements
 10 codes involving 2 elements

COMPOSITE FINGERPRINT INDEX

8	9	9	9	9	9	9	9	9
7	0	0	0	0	0	0	0	0
6	0	10	10	10	10	10	0	10
5	0	29	7	28	29	29	0	23
4	0	4	1	5	5	5	0	0
3	0	0	1	4	3	3	0	1
2	0	9	1	1	0	1	0	8

1	1	2	3	4	5	6	7	8

True positives..

ANX1_HUMAN	ANX1_BOVIN	ANX1_CAVCU	ANX1_RAT
ANX1_RABIT	ANX1_MOUSE	ANX1_COLLI	ANX1_COLLI
ANX1_RODSP			

Subfamily: Codes involving 6 elements

Subfamily True positives..

093446	ANX2_HUMAN	ANX2_CHICK	ANX2_RAT
ANX2_BOVIN	ANX2_MOUSE	ANX2_XENLA	ANX2_XENLA
093444	ANX5_BOVIN		

Subfamily: Codes involving 5 elements

Subfamily True positives..

093447	ANX3_RAT	ANX5_CHICK	ANX6_MOUSE
035639	ANX4_MOUSE	ANX4_HUMAN	ANX4_RAT
ANXA_BOVIN	ANXB_BOVIN	ANX4_PIG	ANX4_BOVIN
ANXA_RABIT	ANX6_HUMAN	ANX4_CANFA	ANXA_HUMAN
ANX6_RAT	ANX5_RAT	ANX3_HUMAN	ANX5_MOUSE
ANXA_MOUSE	ANX5_HUMAN	ANXD_HUMAN	093445
ANX7_HUMAN	ANX7_MOUSE	ANX6_CHICK	ANXX_DROME
ANXD_CANFA			

Subfamily: Codes involving 4 elements

Subfamily True positives..

ANX8_HUMAN	035640	ANXC_HYDAT	ANX5_CYNPY
Q27512			

Subfamily: Codes involving 3 elements

Subfamily True positives..

ANX7_XENLA	Q27473	ANX7_DICDI	059907
----------------------------	------------------------	----------------------------	------------------------

Subfamily: Codes involving 2 elements

Subfamily True positives..

Q27864	081536	081535	076027
Q43863	024131	Q42657	024132
082090	065848		

[Q27864](#)[081536](#)[081535](#)[076027](#)[Q43863](#)[024131](#)[Q42657](#)[024132](#)[082090](#)[065848](#)

NEX1 ANNEXIN - CAENORHABDITIS ELEGANS.

ANNEXIN P34 - LYCOPERSICON ESCULENTUM (TOMATO).

ANNEXIN P35 - LYCOPERSICON ESCULENTUM (TOMATO).

ANNEXIN 31 (ANNEXIN XXXI) - HOMO SAPIENS (HUMAN).

ANNEXIN P33 - ZEA MAYS (MAIZE).

ANNEXIN - NICOTIANA TABACUM (COMMON TOBACCO).

ANNEXIN - CAPSICUM ANNUUM (BELL PEPPER).

ANNEXIN - NICOTIANA TABACUM (COMMON TOBACCO).

FIBER ANNEXIN - GOSSYPIUM HIRSUTUM (UPLAND COTTON).

ANNEXIN - MEDICAGO TRUNCATULA (BARREL MEDIC).

SCAN HISTORY

OWL21_1	2	100	NSINGLE
OWL26_0	1	100	NSINGLE
SPTR37_9f	2	122	NSINGLE

INITIAL MOTIF SETS

ANNEXIN11 Length of motif = 16 Motif number = 1
 Annexin type I motif I - 1

	PCODE	ST	INT
FLKQAWFIENEEQEYV	ANX1_HUMAN	6	6
FLKQARFLENQEYV	ANX1_MOUSE	6	6
FLKQAYFIDNQEYV	ANX1_CAVCU	7	7
FLKQAWFMENLEQECI	ANX1_COLLI	7	7
FLKQACYIEKQEYV	ANX1_RAT	6	6

ANNEXIN12 Length of motif = 23 Motif number = 2
 Annexin type I motif II - 1

	PCODE	ST	INT
MVKGVDDEATIIDILTKRNNAAQRQ	ANX1_HUMAN	55	33
MVKGVDDEATIIDILTKRTNAQRQ	ANX1_MOUSE	55	33
TVKGVDDEATIIDILTKRNNAAQRQ	ANX1_CAVCU	56	33
TAKGVDDEATIIDIMTTRTNAQRQ	ANX1_COLLI	51	28
MVKGVDDEATIIDILTKRTNAQRQ	ANX1_RAT	55	33

ANNEXIN13 Length of motif = 17 Motif number = 3
 Annexin type I motif III - 1

	PCODE	ST	INT
LKKALTGHLEEVVLALL	ANX1_HUMAN	95	17
LRKALTGHLEEVVLALL	ANX1_MOUSE	95	17
LKKALTGHLEEVVLALL	ANX1_CAVCU	96	17
MKRVLKSHLEDVVVALL	ANX1_COLLI	91	17
LKKALTGHLEEVVLALL	ANX1_RAT	95	17

ANNEXIN14 Length of motif = 22 Motif number = 4
 Annexin type I motif IV - 1

	PCODE	ST	INT
LRAAMKGLGTDEDTLIEILASR	ANX1_HUMAN	122	10
LRGAMKGLGTDEDTLIEILTTR	ANX1_MOUSE	122	10
LRAAMKGLGTDEDTLIEILVSR	ANX1_CAVCU	123	10
LRACMKGHGTDEDTLIEILASR	ANX1_COLLI	118	10
LRAAMKGLGTDEDTLIEILTTR	ANX1_RAT	122	10

- **Integra varias otras bases de datos en un solo lugar y provee referencias a otras bases de datos (GO)**
 - **<http://www.ebi.ac.uk/interpro>**
 - **Prosite, PRINTS, Pfam, ProDom, SMART**

InterPro

InterPro Simple Search

You can use this page to search for InterPro, Pfam, PRINTS, Prosite, SWISS-PROT, TrEMBL accession numbers and names, database names, and entry_types. You may combine more than one search term with 'AND', '&', 'OR', '|', 'NOT' and '!'; you may also use wildcarded expressions (eg. *bar**).

Enter search terms here...

Search results for 'human transporter'

Click on the links below to jump to individual InterPro entries.

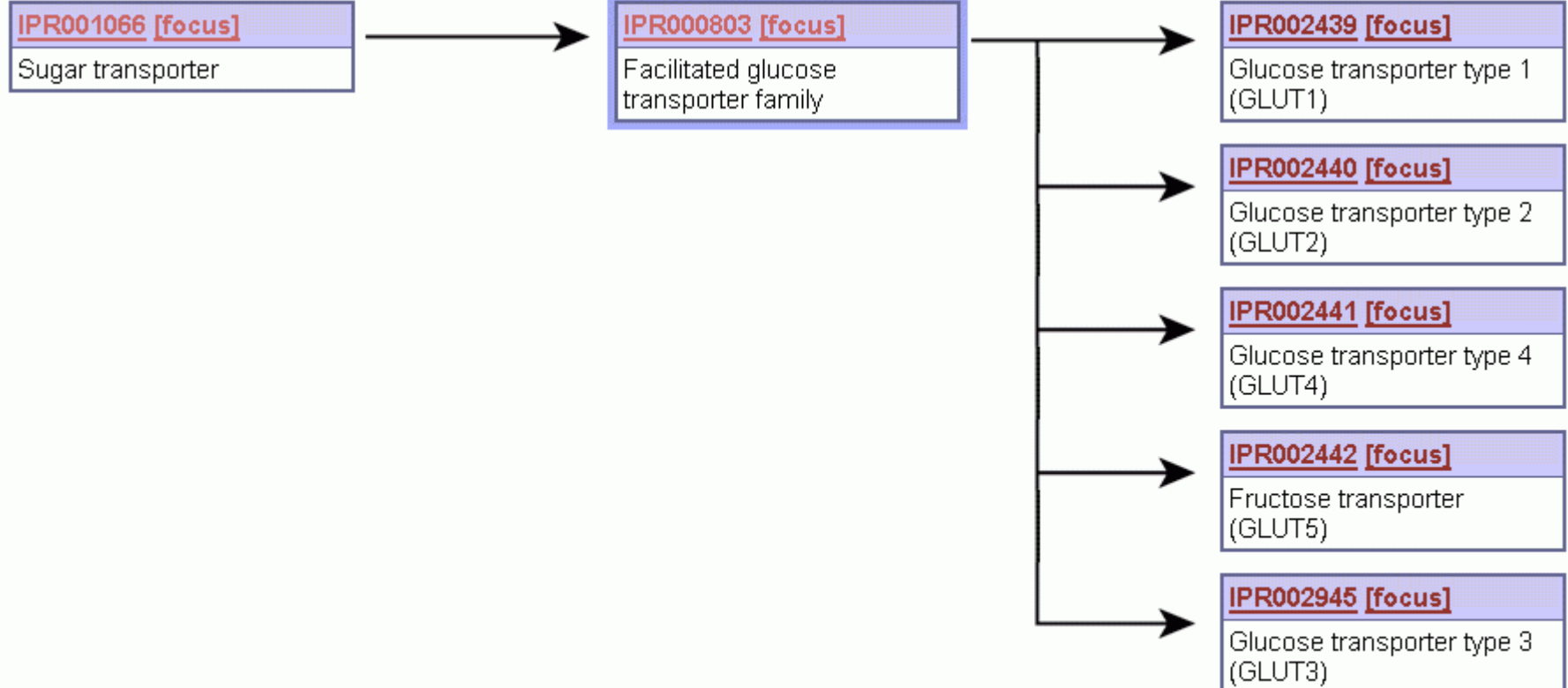
Entry	Entry name
IPR000076	K-Cl co-transporter
IPR000622	K-Cl Co-transporter type 1 (KCC1)
IPR000803	Facilitated glucose transporter family
IPR000849	GipT family of transporters
IPR001066	Sugar transporter
IPR001204	Phosphate transporter family
IPR001902	Sulfate transporter
IPR002259	Delayed-early response protein/equilibrative nucleoside transporter
IPR002293	Permease for amino acids and related compounds, family I
IPR002435	Noradrenaline neurotransmitter transporter
IPR002436	Dopamine neurotransmitter transporter
IPR002437	Serotonin (5-HT) neurotransmitter transporter

InterPro

Tree display for IPR000803

The tree below shows the selected InterPro entry, the path to the root of the tree, the immediate children and the immediate children of the selected entry's parent (i.e. the entry's siblings).

To return to the full entry for this accession number, click [here](#).



InterPro

InterPro - Proteins matching IPR000803

Table Graphical

Grid shows 10aa intervals, first mark at position 0. Move the mouse over a match to see more information in the status line of your browser window.

Item 21-40 of 91

< 1 2 3 4 5 >

Protein	Match Display	
SWISS-PROT GTR2_HUMAN P11168	IPR000803 PR00172	GLUCTRNSPORT
	IPR001066 PS00216	SUGAR_TRANSPORT
	IPR001066 PS00217	SUGAR_TRANSPORT
	IPR001066 PR00171	SUGRTRNSPORT
	IPR001066 PF00083	sugar_tr
	IPR002440 PR01191	GLUCTRSPORT2
SWISS-PROT GTR3_HUMAN P11169	IPR000803 PR00172	GLUCTRNSPORT
	IPR001066 PS00216	SUGAR_TRANSPORT_1
	IPR001066 PS00217	SUGAR_TRANSPORT_2
	IPR001066 PR00171	SUGRTRNSPORT
	IPR001066 PF00083	sugar_tr

Help for : graphic key - Netscape

Graphical match display legend

The table below shows the colour coding used in the graphical match display. The extent of the bars denotes the region on the protein sequence that the selected method [matches](#).

	True	Unknown
PRINTS		
PROSITE pattern		
PROSITE profile		
PFAM		
ProDom		n/a

See also :

InterPro

Help for : table legend - Netscape

Tabular match display legend

The single letter codes after the amino acid ranges in this table denote the status of each individual match. Possible values are shown in the table below :

- T True
- F False Positive
- N False Negative
- P Partial
- ? Unknown

InterPro - Proteins matching IPR001066

Table [Graphical](#)



Item 401-420 of 1177

< [Previous](#) [21](#) [22](#) [23](#) [24](#) [25](#) [Next](#) >

	PS00216	PS00217	PR00171	PF00083
P39637 YWFA_BACSU				19-406 T
P39843 BMR2_BACSU	65-81 T			17-398 T
P39850 CAPA_STAAU		175-200 F		
P39924 HXTC_YEAST	370-387 T	169-194 T	68-78 T 164-183 T 328-338 T 423-444 T 446-458 T	60-521 T
P39932 STL1_YEAST	347-364 T	N		30-488 T
P40441 YIRO_YEAST	263-280 T	62-87 T		2-416 T
P40474 YIM1_YEAST	117-133 F			61-539 T
P40475 YIM0_YEAST	125-141 F			71-547 T
P40862 PROP_SALTY	P	P		
P40885 HXT9_YEAST	373-390 T	172-197 T	72-82 T 167-186 T 331-341 T	64-526 T

MSA: frecuencias de sustitución de aas

- **Un MSA es la base para determinar las frecuencias de sustitución de amino ácidos en un grupo particular de secuencias**
 - **frecuencias de sustitución globales**
 - Se utilizan para generar matrices de scoring:
 - Matrices PAM, BLOSUM, etc
 - Dan puntaje y penalizan por igual los mismos cambios, independientemente del contexto
 - **frecuencias de sustitución sitio por sitio**
 - Position Specific Scoring Matrices (PSSM)
 - Profiles

Bioinformatics. Sequence and Genome analysis. David W Mount,
CSHL Press (2001)

Markov Chains, a visual explanation

<http://setosa.io/blog/2014/07/26/markov-chains/index.html>

**Schneider Lab Home Page (Information Theory for Biology,
Sequence Logos)**

<http://schneider.ncifcrf.gov/>