

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'.

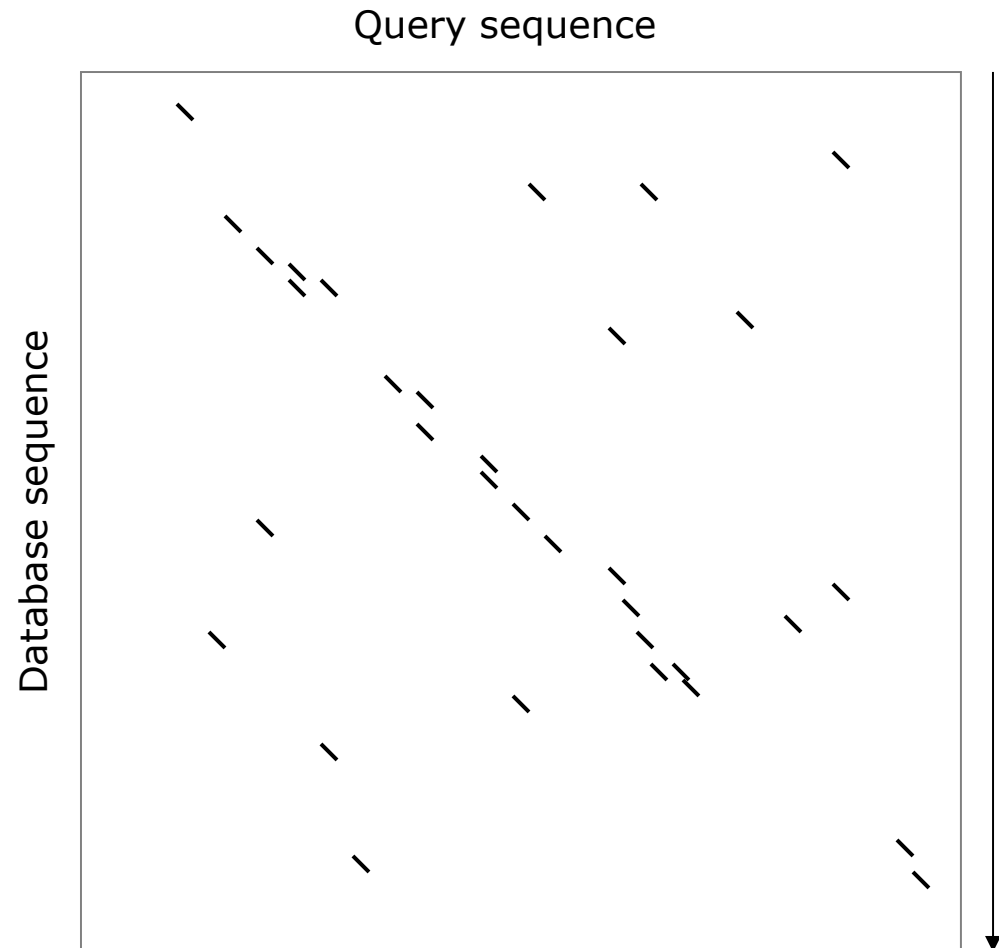
```
> fasta myquery swissprot -ktup 2
```

```
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|NAPÄ_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
```


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 $20^{k_{tup}}$ palabras (proteínas = 20 aminoácidos o letras posibles)

Para $k_{tup}=3$ 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**

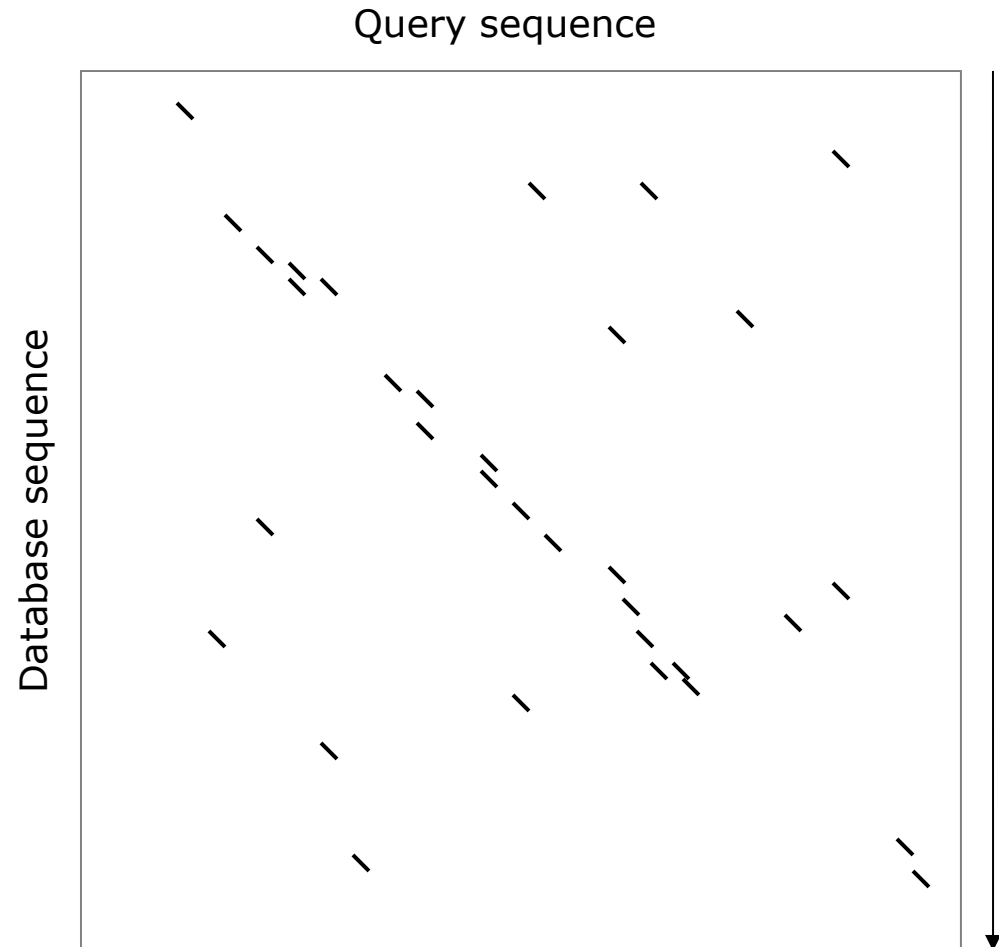
query
sequence

all overlapping
words of size 3

Hashing methods

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

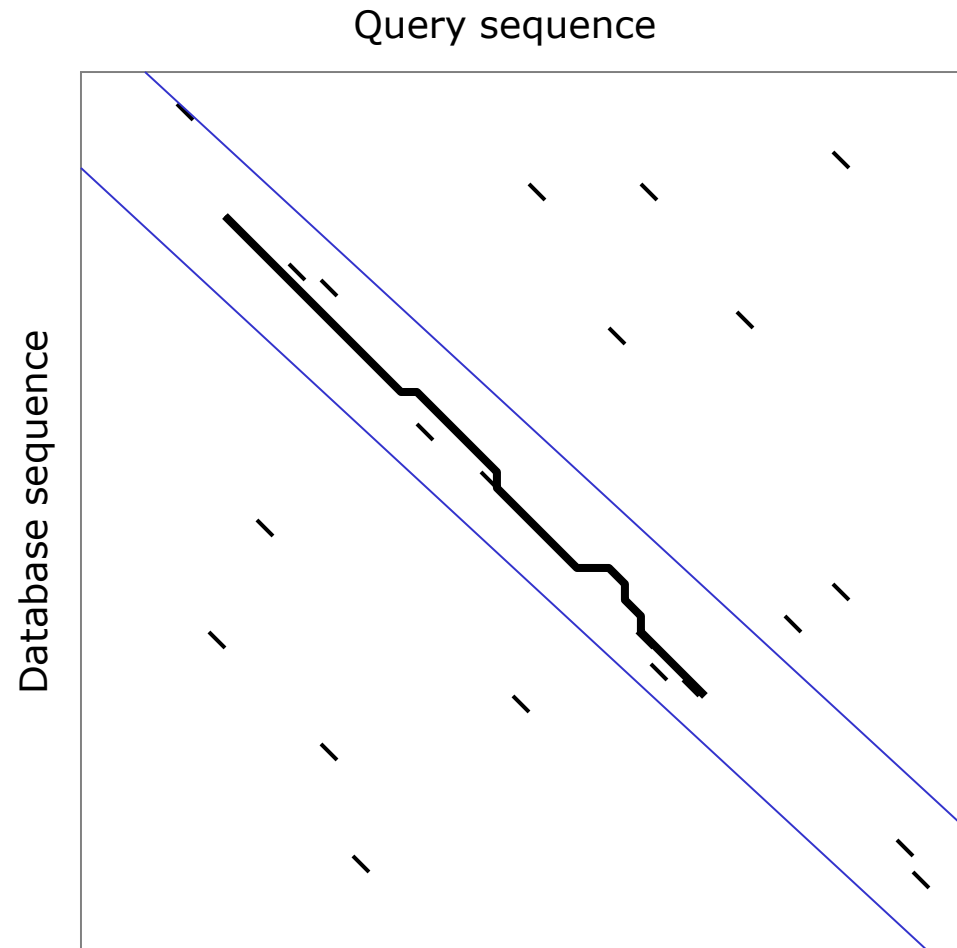
Use word hits to
determine where to search
for alignments
fills the dynamic
programming matrix
in $(N-2)+(M-2)$ operations
instead
of $M \times N$.



Hashing methods

Scan the database,
looking up words in
the dictionary

Use word hits to
determine where to search
for alignments

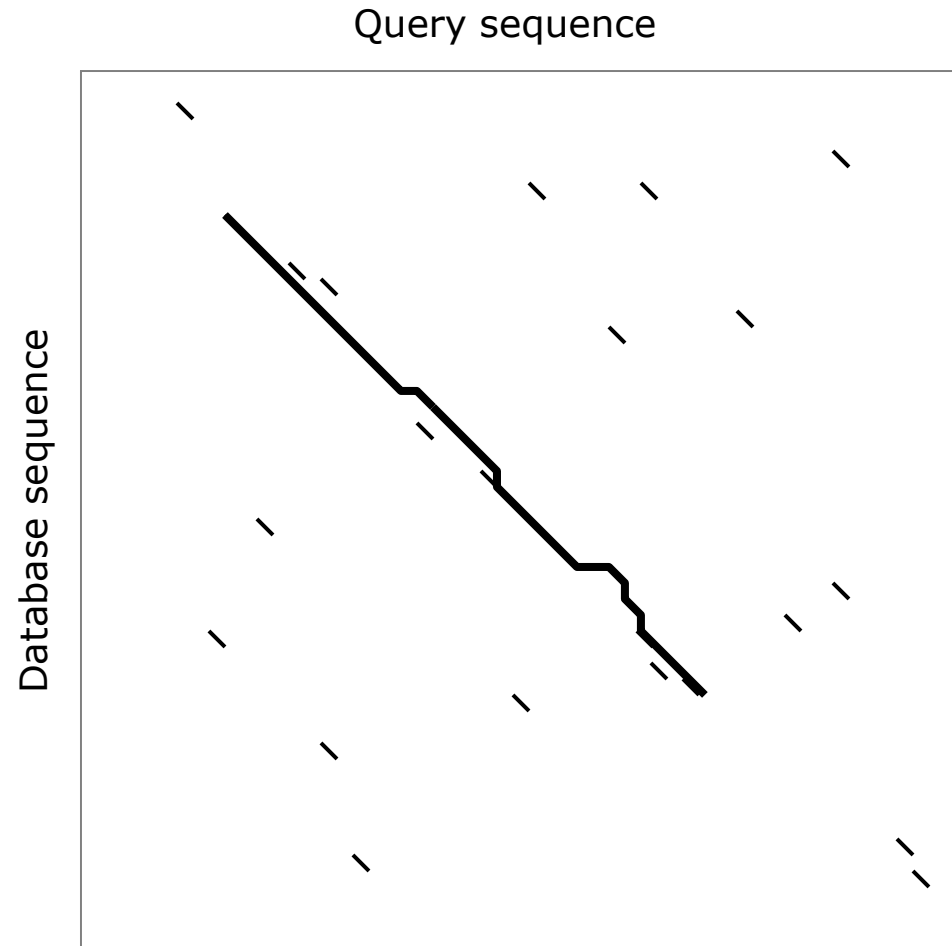


FASTA searches in a band

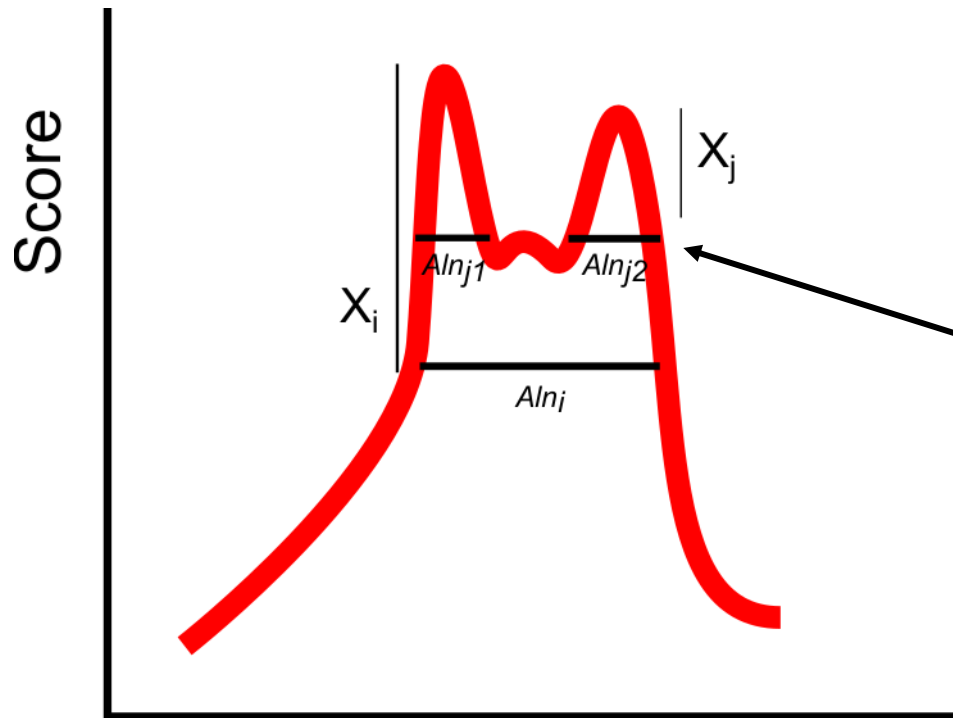
Hashing methods

Scan the database,
looking up words in
the dictionary

Use word hits to
determine where to search
for alignments



un HSP es un “*high scoring pair*”

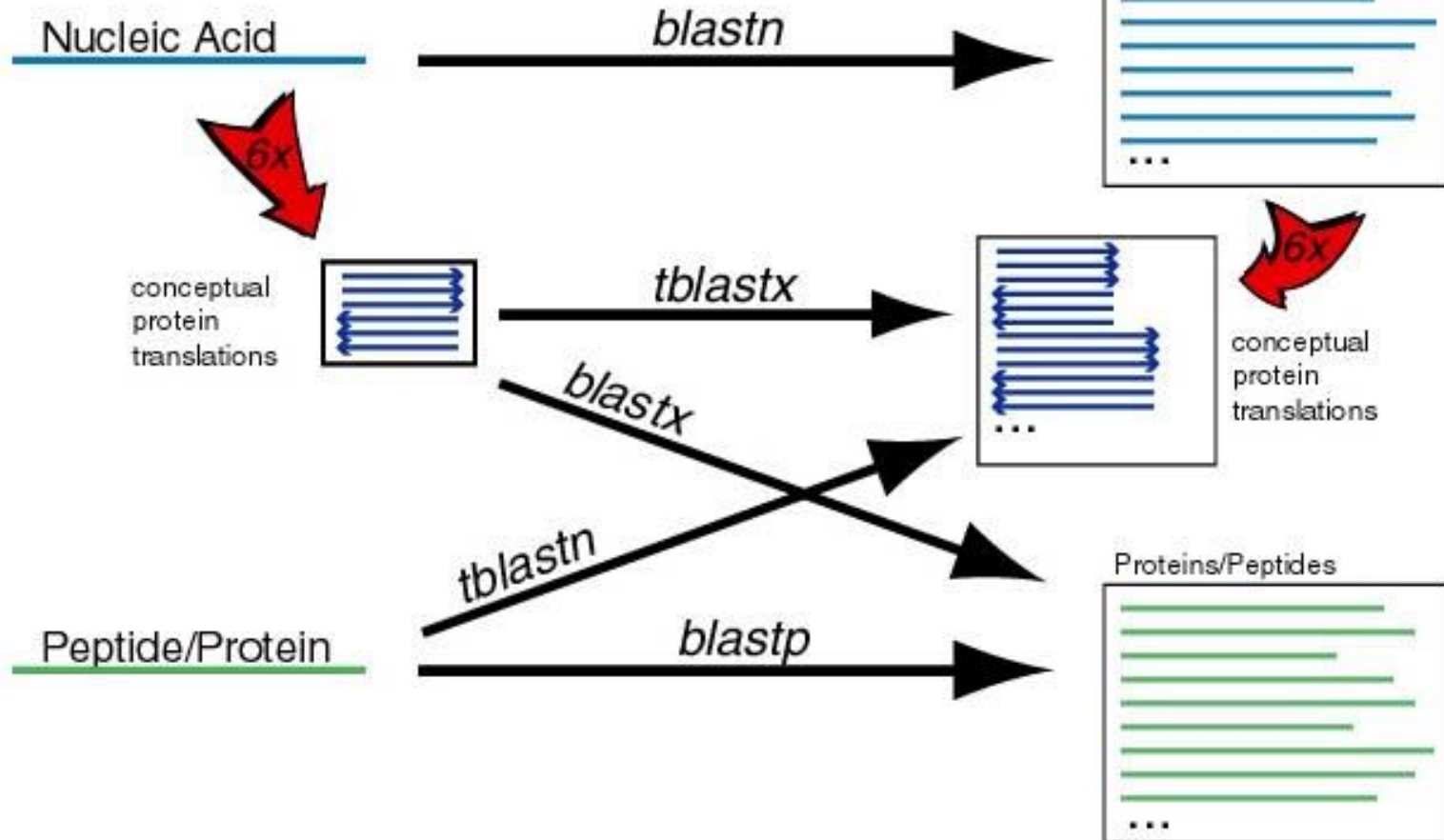


BLAST intenta extender el HSP, siempre que la caída del score sea menos que X (bits). Si lo logra, se repite con el próximo pico.

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**

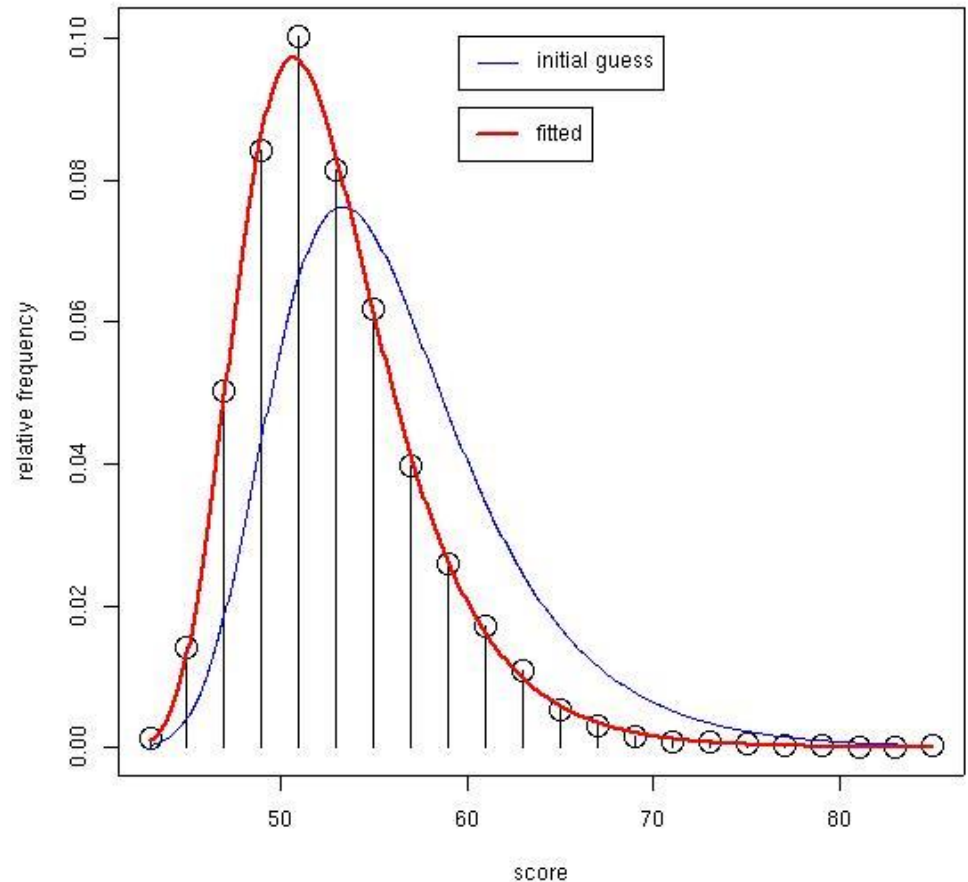
The Gumbel/Extreme value distribution

- 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'.

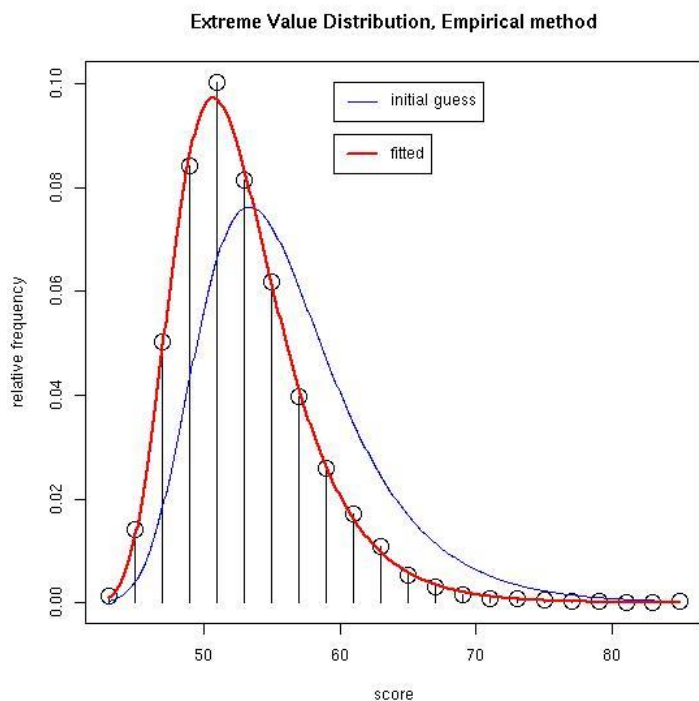
E-value: Número de alineamientos con un score = S que se espera encontrar si la base de datos es una colección de letras al azar.

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

$$E(S > x) = p(S > x) D$$

- 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)**

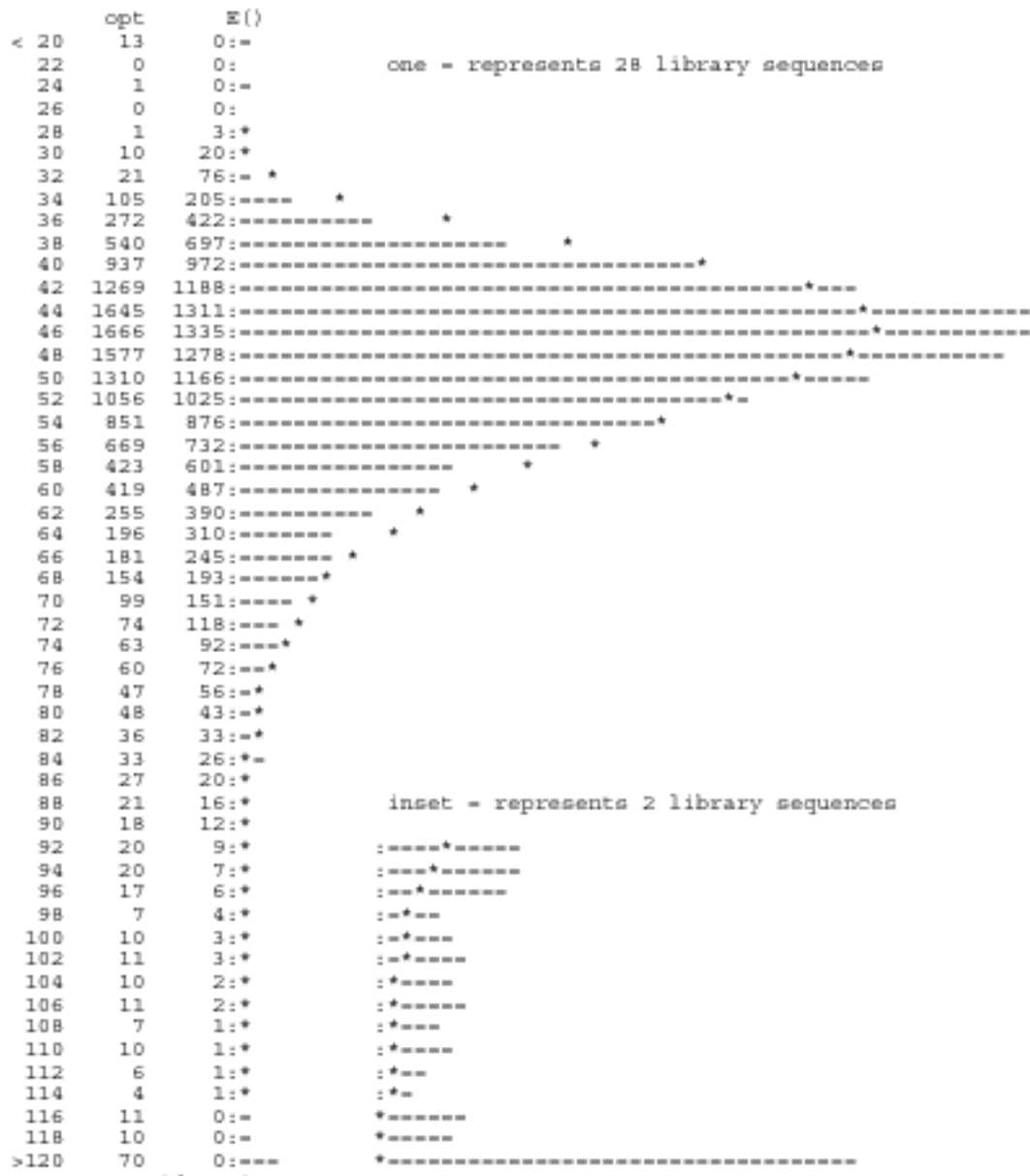
- **En ciertos casos, la estadística de los alineamientos falla**
 - **Lo que falla son las suposiciones que hicimos para llegar al modelo estadístico que describe - en este caso - la distribución de scores entre secuencias no relacionadas**
- **En general se obtienen estimaciones incorrectas de E cuando**
 - **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:      one = represents 22 library sequences
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:=
```

Mirar el histograma de scores esperados y observados
Mirar el E de la secuencia no relacionada con mayor score

Evaluando la estadística (cont)



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

/note= Phosphoserine. (ECO:0000250|UniProtKB:P45481).

ORIGIN

```
1 maenlldgpp nprkrklssp gfsandstdf gslfdlendl pdelipngge lgllnsgnlv
61 pdaaskhkhql sellrgsgs sinpgignvs asspvqqglg gqaqqpnasa nmaslsamgk
121 splsqgdssa psllpqaast sgtpaasqa lnpqaqkqvg latsspatsq tpggicmnan
181 fnqthpglln snsghslinq asqgqaavmn aslaaaarar gaampvptpa mggassyvla
```

```
241 etltqvspqm tghaglnaq aggma
301 qsmvnslptf ptdikntsvt nvpnm
361 llhahkcqrr eqangevrac slphc
421 nctrhdcpvc lplknasdkr nqqti
481 mqrayaalgf pymnqpqtql qpqvpe
541 qppnlisesa lptslgatnp lmdng
601 rshlvhklvq aifptdpaa lkdrri
661 iqkeleekrr srlhkqgilg nqpal
```

```
721 mnsfnpmisg nvqlpqapmg praaspmnhs vqmnsmsgsvp gmaispsrmp qppnmgaht
781 nmmaqapaaq sqflpqnqfp sssgamsvgm gqppaqtgvs qgqvpgaalp nplnmlgpqa
841 sqlpcppvtq splhptppa staagmpslq httpgmtpp qpaaptqpst pvsssgqtpt
901 ptpgsvpsat qtqstptvqa aaqaqvtpqp qtpvqppsva tqssqqqpt pvhaqppgtp
961 lsqaaasidn rvptpssvas aetnsqqqpg dvpvlemkte tqaedtepd geskgeprse
1021 mmeedlqgas qvkeetdiae qksepmevde kkpevkvevk eeeessngt asqstpsqp
1081 rkkifkpeel rqalmptlea lyrqdpeslp frqvpdpql gipdyfdivk npmldstikr
1141 kldtgqyqep wqyvddvwl mfnawlynrk tsrvykfcsk laevfeqid pvmqslgycc
1201 grkyefspqt lccyqkqlct iprdaayysy qnrhfcck fteiagenvt lgddpsqqpt
1261 tiskdqfek kndtldpepf vdckecgrkm hqicvlhydi iwpsgfvcn clkktgrprk
1321 enkfsakrlq ttrlgnhled rvnkflrrqn hpeagevfv vvasdktve vkpgmksrfv
1381 dsqsesesfp yrtkalrafe eidgvdcff gmhvqeygsd cppntrrvy isylsiahff
1441 rprclrtavy heiligyley vkklgyvtgh iwacppsegd dyifhchppd qkipkprlq
1501 ewykkmlcka faerihdyk difkqatedr ltsakelpyf egdfwpvle esikeleqee
1561 eerkkeesta asetegsqg dsknakkkn kktknkssi srankkpsm pnsndlsqk
1621 lyatmekhke vffvihlhag pvintlppiv dpdplscdl mdgrdafll ardkhwefss
1681 lrrskwstlc mlvelhtqgq drfvytcnec khvetrwhc tvcedydlci ncyntkshah
1741 kmvkwglgld degssqgepq skspqesrll siqrciqslv hacqcrnanc slpscqkmkr
1801 vvqhtkgckr ktnggcpvck qlialccyha khcqnkcqv pfclnikhkl rqqqihrlq
1861 qaqlmrrrma tmntrnvpqq slpsptsapp gtptqqpstp qtpppaqqp pspvmspag
1921 fpsvartqpp ttvstgkpts qvpappppaq pppaaveaar qiareaqqq hlyrvninns
1981 mppgrtgmt pgsqmapvsl nvprpnqvsg pvmpsmppgq wqqaplpqq pmpglprpvi
2041 smqaqaavag prmpsvqppr sispsalqdl lrtlkspssp qqqqqvlnil ksnplmaaf
2101 ikqrtakyva nqpgmqppg lqsqpgmqpp pgmhqqpslq nlnamqagvp rpgvppqqqa
2161 mgglnpqgga lnimnpghnp nmasmnpqyr emlrrqllq qqqqqqqqqq qqqqqqsag
2221 maggmaghgq fqqpqqpggy ppamqqqqrm qqhlplqgss mgqmaaamgq lgqmqppglg
2281 adstoniqaq laarilaqaq apomsqaahm lsqaaashl paaiaatsls
```

ORIGIN

```
1 mmsylkppy avnglsltts gmdllhpsvg ypatprkqrr erttftraq l dvlealfakt
61 rydpdifmree valkinlpes rvqwvfkrr akcrqqqqqq qnggqkvrv akkksspare
121 vssesgtsgq ftpsstsyp tiasssapvs iwspasispl sdplstssc mqrssypmyt
181 qasgysqgya gtsyfggmd cgsyltpmhh qlpgpgatls pmgtnavtsh lnqspaslst
241 qygasslgf nsttdcldyk dqtaswklf nadcldykq tsswkfql
```

//

.1977519J.

```
lmd intlnggssd tadkirihak nfeaalfaks
vta aaannnikpv eqhhinnlkn sgnsanmrv
qqq qqqqqqqrr qltpqqqlv nqmkvapik
ltp qdmeaakevy kihqqllfka rlqqqqaq
mqp pnsannnpl qqqssqntvp nvlqnqif
mte pvkqsfirky inqalrkiq alrdvknenn
nnn dtiatsatpn aaafsqqna ssklyqmqq
qaq aqaqqaqqaq aqaqqaqqaq aqaqqaqqaq
akd vevikqlsld asktnlrld vtlnlsneek
tkn enflkevflq rifvkeilek caegifvkl
lrq qqqmannngn pgttstgnnn niatqqnmq
qqq qqqqqqhiyp sstpgvanys amanapgnni
aat pslnktingk vngrtksnti pvtsipstnk
nps plktqtkngt 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--LCGSRLTTTLLAVCRNQLCTGLTAFKRSADQSY----- 66

Query 71  GPGAGSLQPLALEGSLQKRG-IVEQCCTSISLQSLYQLENYC 109
A + + L QKRG I +CC CS L+ +C
Sbjct 67  ---APTTRDLFHIHQQKRGGIATECCEKCSFAYLKTF 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--LCGSRLTTTLLAVCRNQLCTGLTAFKRSADQ-----S 65

Query 70  GGPGAGSLQPLALEGSLQKRGIVEQCCTSISLQSLYQLENYC 109
P L + ++ GI +CC CS L+ +C
Sbjct 66  YAPTTRDL--FHIHQQKRGGIATECCEKCSFAYLKTF 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 ISGAILFEETLFQKNEAGVPMVNLLHNENIIPGIKVDKGLVNIPTDDEE--KSTQGLDGLAERCKEYKAGA 147
I GAILFE+T+ K ++ L + ++P +K+DKGL ++ + K L L +R E + G
67 ILGAILFEQTMSKIDGKYTADFLWEEKVLPFLKIDKGLNDLDADGVQTMKPNPTLADLLKRANERHIFG- 137

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

(b)

```
78 ISGAILFEETLFQKNEAGVPMVNLLHNENIIPGIKVDKGLVNIPTDDEE--KSTQGLDGLAERCKEYKAGA 147
I GAILFE+T+ K + L + + ++P +K+DKGL ++ + + K L L +R+ E + G
67 ILGAILFEQTMSKIDGKYTADFLWEEKVLPFLKIDKGLNDLDADGVQTMKPNPTLADLLKRANERHIFG- 137

148 RFAKWRTVLVIDTAKGKPTDLS-IHETAWGLARYASICQONRLVPIVEPEILADGPHSIEVCAVVTQKVLSC 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 HLNALPETS NVMLKLT L--PTVENLYEEFTKH-----PRVVRVVALSGGYSREKAN 248
```

(c)

```
78 ISGAILFEETLFQKNEAGVPMVNLLHNENIIPGIKVDKGLVNIPTDDEE--KSTQGLDGLAERCKEYKAGA 147
I GAILFE+T+ K + L + + ++P +K+DKGL ++ + + K + L L +R+ E + G
67 ILGAILFEQTMSKIDGKYTADFLWEEKVLPFLKIDKGLNDLDADGVQTMKPNPTLADLLKRANERHIFG- 137

148 RFAKWRTVLVIDTAKGKPTDLS-IHETAWGLARYASICQONRLVPIVEPEILADGPHSIEVCAVVTQKVLSC 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 LNALPETS NVMLKLT L--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