Identificación de homólogos lejanos mediante PSI-BLAST

La búsquedas de secuencias distantes en bases de datos mediante **matrices de ponderación sitio específicas** (también conocidas como **perfiles** o **motivos**) son generalmente más adecuadas para la identificación de homólogos con bajo nivel de identidad que el BLASTP estándar

PSI-BLAST (**Position-Specific Iterated BLAST**) es una modificación de BLASTP que permite la búsqueda de homólogos mediante **perfiles generados automáticamente a partir de alineamientos múltiples derivados de los HSPs encontrados por BLASTP**.

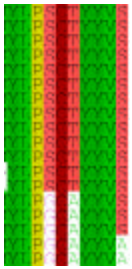
- Pasos que sigue el algoritmo de PSI-BLAST

1. Búsqueda de homólogos de una sec. problema mediante BLASTP
2. Construcción de un aln. múltiple a partir de los HSPs y construcción de un perfil
3. El programa compara el perfil construido con la base de datos
4. PSI-BLAST determina la significancia estadística de los alns. locales encontrados
5. PSI-BLAST puede repetir o iterar los pasos a partir del 2. para construir perfiles cada vez más específicos con las secuencias nuevas encontradas en cada iteración hasta llegar a la convergencia

Identificación de homólogos lejanos mediante PSI-BLAST

- **matrices de ponderación sitio específicas (Position Specific Scoring Matrices PSSMs)**

Se construyen usando algoritmos de cadenas ocultas de Markov (**HMMs**). En esencia, para un alineamiento múltiple se consideran tanto las posiciones como las frecuencias de los estados de caracter observados para cada sitio. Residuos muy conservados en una determinada posición reciben un score positivo muy alto, mientras que los raros en dicha posición reciben un score alto negativo. Resíduos que ocupan posiciones muy variables reciben scores próximos a cero.



```
          12345678910
          A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
 1 Y     -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1
 2 L     -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1
 3 P     -1 -2 -2 -2 -3 -2 -1 -2 -2 -3 -3 -1 -3 -4  8 -1 -1 -4 -3 -3
 4 S      1 -1  0 -1 -1  0  0 -1 -1 -3 -3  0 -2 -3 -1  5  1 -3 -2 -2
 5 C     -1 -4 -3 -4  9 -3 -4 -3 -3 -2 -2 -3 -2 -3 -3 -1 -1 -3 -3 -1
 6 T      0 -1  0 -1 -1 -1 -1 -1 -2 -2 -3 -1 -2 -3 -1  4  3 -3 -2 -2
 7 Y     -2 -3 -3 -4 -3 -2 -3 -4  1 -1 -1 -3 -1  5 -4 -2 -2  1  7 -2
 8 Y     -1 -1 -1 -1 -2  0 -1 -2  6 -2 -1 -1 -1  1 -1 -1 -1  0  5 -2
 9 V     -1 -2 -2 -2 -1 -2 -2 -2 -2  1  2 -2  0 -1 -2 -2 -1 -2 -1  4
10 S     -1 -1 -1 -1 -3  3  3 -2 -1 -2  1  0 -1 -2 -2  2 -1 -3 -2 -2
```

Ejemplo de una PSSM calculada para los 10 primeros resíduos de un alineamiento múltiple de proteínas HoxA de eucariontes. Sólo se muestra una pequeña parte de las secuencias incluídas en el alineamiento múltiple usado para calcular la PSSM

Identificación de homólogos lejanos mediante PSI-BLAST

Identificación de homólogos lejanos mediante PSI-BLAST

Identificación de homólogos lejanos mediante PSI-BLAST

Identificación de homólogos lejanos mediante PSI-BLAST

• **Aspectos a cuidar al calcular PSSMs**

1.- Hay que evitar a toda costa incluir secuencias no homólogas. Revisar alineamientos pareados, estructura de dominios y no fiarse de las anotaciones. Muchas secuencias están mal anotadas !!!

Utilizar:

http://www.ncbi.nlm.nih.gov/COG/
http://psort.hgc.jp/
http://www.predictprotein.org/newwebsite/
http://www.ch.embnet.org/software/TMPRED_form.html
http://www.expasy.org/
...

para caracterizar a las proteínas dudosas ...

2.- Eliminar regiones de baja complejidad.

Usar SEG y COILS

http://www.ch.embnet.org/software/COILS_form.html

PRÁCTICAS: aprendiendo a usar PSI-BLAST para identificar homólogos lejanos

1) Descarga la secuencia Q57997 y haz un análisis de PSI-BLAST. Preguntas:
   - Qué tipo de función podría tener esta proteína?
   - Cuantos homólogos encontraste en la primera búsqueda (BLASTP)
   - Cuantos ciclos o iteraciones tuviste que correr hasta la convergencia? Cuantos homólogos pescaste?

2) Compara estos resultados con el análisis descrito en el tutorial de PSI-BLAST que encontrarás en la página del NCBI bajo:
   http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html

3) Ve a la página de nuestro curso y haz los ejercicios propuestos que encontrarás en el directorio Ejercicios/BLAST

Consejos finales para el uso eficiente de BLAST

0. Antes de iniciar búsquedas con BLAST, hay que **escanear las secs.** para detectar la presencia de múltiples dominios, reg. repetitivas, motivos y péptidos señal usando las herramientas o servidores apropiados (**SMART, PROSITE, PFAM, CDD, PSORT** ...)
1. Para búsquedas de secuencias homólogas distantes **usa AAs y PSI-BLAST** siempre que sea posible.
2. **PSSMs**. Usa todos los criterios adicionales que consideres relevantes para inferir la homología de manera certera. No te fíes de las anotaciones, las hay erróneas. También conviene ser crítico con las proteínas hipotéticas, puesto que su existencia no se ha demostrado experimentalmente y con frecuencia presentan extremos N terminales más largos que los de las proteínas de verdad (problema de predecir adecuadamente el inicio de traducción).
3. Ajusta el valor de los parámetros de búsqueda de manera adecuada al problema a resolver. **El valor de los parámetros determina lo que puedes encontrar**. Así por ejemplo búsquedas con NCBI-BLASTN con valores por defecto de match (+1) y mismatch (-3) tienen una frecuencia diana de 99% de identidad. No busques genes de humano y nemátodo con NCBI-BLASTN...
4. Haz **controles**, especialmente cuando se trate de similitudes en la **zona de penumbra** Así por ejemplo puedes hacer un "**barajado" de la secuencia problema** a mano o mejor aún, usando un sencillo script de Perl. Si después de barajar los caracteres de tu secuencia sigues encontrando hits similares en la zona de penumbra, el parecido se debe simplemente a un sesgo composicional compartido entre ambas secs. y no a homología

URLs de algunas de las principales bases de datos de secuencias (DNA, Prot.), familias/dominios/motivos de proteínas y estructuras

Blocks and Blocks+ : http://blocks.fhcrc.org/
DBJ : http://www.ddbj.nig.ac.jp/
EMBL : http://www.ebi.ac.uk/embl/
Entrez : http://www.ncbi.nlm.nih.gov/Entrez/
GenBank : http://www.ncbi.nlm.nih.gov/Genbank/
InterPro : http://www.ebi.ac.uk/interpro/
MEDLINE : http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
PDB : http://www.rcsb.org/pdb/
PIR : http://www-nbrf.georgetown.edu/
Pfam : http://www.sanger.ac.uk/Pfam/
PRINTS : http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html
ProDom : http://protein.toulouse.inra.fr/prodom.html
PROSITE :http://www.expasy.ch/prosite/prosite.html
SRS "mother" server :http://srs.ebi.ac.uk/
SWISS-PROT and TrEMBL at EBI : http://www.ebi.ac.uk/swissprot/