

Identificación de homólogos lejanos mediante PSI-BLAST

Las 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 carácter 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. Residuos que ocupan posiciones muy variables reciben scores próximos a cero.

	1	2	3	4	5	6	7	8	9	10	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	2	7	-1										
2 L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-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	-2	0	-1	-2	6	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	5	-2										
9 V	-1	-2	-2	-1	-2	-2	-2	-2	1	2	-2	0	-1	-2	-2	-1	-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 residuos de un alineamiento múltiple de proteínas HoxA de eucariontes. Sólo se muestra una pequeña parte de las secuencias incluidas en el alineamiento múltiple usado para calcular la PSSM

Identificación de homólogos lejanos mediante PSI-BLAST



Identificación de homólogos lejanos mediante PSI-BLAST

Sequence	Score (Bits)	E-Value
gi186399076 cwf1P_470999.11 acid tolerance and virulence pro...	796	0.0
gi124265566 gh1AAH52119.11 kva precursor [Shigella flexneri]	475	2e-132
gi13791017 gh1AA45509.11 agrobacterium chromosomal virulence...	465	7e-126
gi1056293 gh1AA31937.11 acvB-virulence gene acvB product [Ag...	451	2e-125
gi13010455 gh1AA31142.11 acvB virulence protein B [Stacthob...	435	2e-120
gi1315211 gh1AA03396.11 positive membrane protein [Agrobacter...	379	1e-101
gi1389316 pepf120342288 virulence gene	379	1e-101
gi181927131 gh1AA031676.11 virulence protein [Rhodospirillum ...	339	9e-89
gi13824314 cwf1P_CAC48261.11 COMBEPED HYPOTHETICAL PROTEIN [B...	317	7e-61
gi13895142 cwf1P_MP_530799.11 wip3 [Agrobacterium tumefaciens]...	313	2e-56
gi186186601 cwf1P_665465.11 type IV secretory pathway VirJ c...	113	3e-29
gi181924681 cwf1P_00881098.11 similar to Type IV secretory p...	113	3e-28
gi178931681 cwf1P_00881081.11 similar to Type IV secretory p...	113	3e-28
gi178988081 cwf1P_00881079.11 similar to Type IV secretory p...	113	4e-28
gi168544001 cwf1P_00881081.11 similar to Type IV secretory p...	110	8e-27
gi1301521101 gh1AA31183.11 VirJ [Agrobacterium tumefaciens]	97.0	0.003

Identificación de homólogos lejanos mediante PSI-BLAST

Accession	Description	E-value
gi168544001 ref PF_00592601.1	similar to Type IV secretory p...	3e-71
gi150754131 emb CAC45745.1	HYPOTHETICAL TRANSDUCIBLE PROTEI...	3e-62
gi167135294 db AA024026.1	Type IV secretory pathway VirD com...	1.0E-22
gi164935067 ref PF_01038507.1	virulence protein [Parvulancul...	6e-14
gi130533106 db AA011991.1	VirD [Agrobacterium tumefaciens]	5e-05
gi131921191 db AA017103.1	conserved hypothetical protein [C...	5e-04
gi1071551291 db AA021603.1	Type IV secretory pathway VirD com...	1.0E-46
gi164795067 ref PF_01038507.1	virulence protein [Parvulancul...	1.7E-42
gi130533106 db AA011991.1	VirD [Agrobacterium tumefaciens]	4e-06
gi131921191 db AA017103.1	conserved hypothetical protein [C...	5e-06
gi113333061 db AA027892.1	Predicted hydrolases or scyttran...	5e-03
gi118871897 ref NP_794516.1	hypothetical protein P59704761 [...]	0.000
gi17291851 db AA017564.1	Alpha/beta hydrolase fold [Pseudom...	0.004
gi171231491 db AA024502.1	conserved hypothetical protein Pse...	0.005
gi130533106 db AA011991.1	VirD [Agrobacterium tumefaciens]	6e-08
gi166043507 ref PF_1173386.1	hypothetical protein Ppr_4300 [...]	1e-08
gi130199187 emb CA048012.1	Punative hydrolase [Corynebacteri...	1e-07
gi166043506 ref PF_1173387.1	hypothetical protein Ppr_1052 [...]	3e-07
gi164311643 ref PF_00499998.1	COG1073: Hydrolases of the alp...	1e-07
gi139947452 db AA025069.1	hypothetical protein PA1690 [Pseud...	1e-07

Identificación de homólogos lejanos mediante PSI-BLAST

• Aspectos a cuidar al calcular PSSMs

- 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 ...

- 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

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

- Antes de iniciar búsquedas con BLAST, hay que **escanear las secs.** para detectar la presencia de múltiples dominios, rep. repetitivas, motivos y péptidos señal usando las herramientas o servidores apropiados (**SMART, PROSITE, PFAM, CDD, PSORT ...**)
- Para búsquedas de secuencias homólogas distantes **usa AAs y PSI-BLAST** siempre que sea posible.
- PSSMs.** Usa todos los criterios adicionales que consideres relevantes para inferir la homología de manera certera. No te fies 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).
- 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...
- 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/>