Protocol per a la detecció de selenoproteïnes en un genoma

1. Obtenció de la seqüència genòmica de l'espècie problema

Primer de tot buscarem la nostra seqüència en una de les bases de dades (les més usades són l'Ensembl, NCBI i Santa Cruz). Tot i així hi ha seqüències de microorganismes que no es troben disponibles en aquestes bases de dades, pel que les haurem de buscar en altres de menys utilitzades. Un cop obtingudes les seqüències les guardarem en format FASTA.

2. Caracterització de l'espècie i antecedents d'estudis amb selenoproteïnes en aquesta

Buscarem informació sobre l'espècie d'estudi i mirarem si hi ha estudis previs sobre selenoproteïnes per tal d'orientar la nostra investigació.

Cerca de selenoproteïnes ja conegudes en el genoma:

3. Cerca de selenoproteïnes en el SelenoDB

Entrarem en la base de dades SelenoDB (**http://www.selenodb.org**) , on sabem que hi ha disponible la seqüència tan genòmica com aminoacídica de totes les selenoproteïnes eucariotes fins ara descobertes. Aquesta pàgina web també proporciona informació sobre els elements Secis i molècules relacionades. A més, separa les proteïnes segons l'espècie, i dins de cada espècie les classifica segons si són selenoproteïnes, homòlegs amb cisteïna o altra maquinària molecular.

Com que l'espècie humana és la que està més estudiada i la que té més selenoproteïnes caracteritzades, utilitzarem aquestes seqüències.

Ens interessa tenir aquesta informació per a buscar seqüències homòlogues en el nostre genoma. En aquest cas agafarem la seqüència aminoacídica per a posteriorment fer el Blast que considerem més adequat per aquest cas.

4. Comparació de les seqüències aminoacídiques obtingudes amb el genoma d'estudi: TBLASTN

Un cop tenim les seqüències d'aminoàcids de les selenoproteïnes humanes, usarem el programa BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool), que ens permet buscar les seqüències homòlogues al nostre organisme. Hi ha diferents tipus de BLAST, depenent de què es vulgui alinear. En aquest cas utilitzarem TBLASTN, que ens compara un *query* de proteïna contra una seqüència de nucleòtids de la base de dades. Considerem que és el més útil perquè evitem falsos missmatches produïts pel biaix en l'ús de codons.

De tots els paràmetres del BLAST que es poden modificar a l'hora de fer alineament, n'hi ha un que ens interessa especialment: l'*E-value.* Aquest és el nombre esperat d'alineaments que podem obtenir amb un *score* igual o superior per atzar en un alineament múltiple. D'aquesta manera, com menor sigui l'*E-value,* més significant és l'alineament. En el nostre cas, ens interessa una gran significància i, per tant, delimitem un *E-value* de 10^{-10}.

El BLAST el podrem utilitzar des de la pàgina web, des de la finestra terminal del Linux o amb la blastmachine (ens permetria córrer totes les seqüències d'un sol cop) segons ens convingui.

5. Selecció dels alineaments significatius i extracció del fragment de la seqüència genòmica d'interès.

De tots els alineaments ens quedarem amb aquells que tinguin un E-value significatiu i que en l'alineament coincideixi la selenocisteïna amb un codó STOP o una cisteïna del nostre genoma.

A partir d'aquí utilitzarem la seqüència genòmica dels alineaments que considerem vàlids i l'allargarem uns 3000pb per tal d'englobar l'element Secis. Creiem que la manera més pràctica de fer-ho és a través de la finestra terminal del Linux (tal i com ens han explicat a pràctiques).

6. Deducció de l'estructura exònica del nostre fragment de genoma: Genewise

Per a poder conèixer l'estructura exònica haurem de comparar el nostre fragment de genoma amb la selenoproteïna que estem estudiant. Això ho podem fer amb el programa Genewise (http://www.ebi.ac.uk/Wise2/index.html) , el qual dedueix els exons i introns de la seqüència genòmica. A partir d'aquí podrem veure si realment els codons que codifiquen selenocisteïna o cisteïna es troben dins un exó, la qual cosa ens indicaria que estem davant una selenoproteïna o un homòleg d'aquesta.

7. Busca d'elements Secis en les selenoproteïnes trobades

Un cop trobades les selenoproteïnes utilitzarem el programa SECISsearch (http://genome.unl.edu/SECISearch.html) per a trobar l'element SECIS en la regió 3' de la seqüència (recordem que abans l'hem allargat amb aquest propòsit). Aquest programa ens permetrà no tan sols localitzar dominis potencials d'elements SECIS sinó que també ens informarà de l'estabilitat termodinàmica d'aquests.

Cerca de noves selenoproteïnes en el genoma:

Tan si no hem trobat selenoproteïnes homòlogues en el nostre genoma com si volem extendre més la nostra cerca, podem buscar la presència de noves selenoproteïnes en el nostre genoma seguint els següents passos:

8. Cerca d'elements SECIS

Podem intentar determinar els elements SECIS que es troben en tot el nostre genoma gràcies al programa *SECISsearch*. També podríem emprar el programa *PatScan* (**http://www-unix.mcs.anl.gov/compbio/PatScan/HTML/scanner.html**) per a aquesta cerca, però en aquest cas hauriem d'utilitzar també el programa RNAfold (**http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi**) per a l'estudi de l'estabilitat termodinàmica.

9. Limitació dels SECIS a regions codificants per possibles selenoproteïnes

Com que del punt anterior esperaríem obtenir un nombre considerable d'elements SECIS potencials hem de limitar la nostra cerca condicionant aquests elements SECIS a la presència d'una seqüència codificant prèvia amb un codó per a selenocisteïna. Amb el programa *geneid* **(http://genome.imim.es/geneid.html)** es poden determinar aquestes seqüències. Hem de tenir en compte que haurem d'utilitzar una versió modificada d'aquest programa que tingui en compte que un codó STOP TGA pot ser codificant.

10. Anàlisis posteriors necessaris

Del procés anteriorment descrit ja obtindríem un nombre bastant limitat de selenoproteïnes. És en aquest punt on cal comparar els possibles candidats amb altres gens ja coneguts i descartar aquelles seqüències que presentin incompatibilitats amb les estructures generals dels ESTs. A més a més cal tenir en compte la viabilitat de les estructures secundàries de les potencials selenoproteïnes predites. Amb tots els passos anteriors esperaríem una reducció encara més gran de les selenoproteïnes candidates. Per finalitzar ens faria falta comprovar que la seqüència de les selenoproteïnes que hem trobat no és cap de les que ja havien estat predites anteriorment.