Bioinformatics MSc Project: Pipeline for annotating short open reading frames

The project will center on developing a bioinformatics pipeline for functionally annotating peptides encoded by short open reading frames (sPEPs), integrating published mass spectrometric datasets and motif discovery to in silico predict the function of each sPEP candidate.

Polypeptides encoded by short open reading frames (sPEPs) are translated abundantly in eukaryotic cells, yet their functions remain mysterious. Many short open reading frames (sORFs) are translated in addition to the major protein products from mammalian mRNAs, and also from long non-coding RNA. sORFs range from a few to ~100 codons and thus have typically not been classified as protein-coding. Diverse experimental or bioinformatic approaches provide evidence for the existence of a large repertoir of sPEPs in human cells. 50% of the mammalian mRNAs contain at least one sORF, albeit only a small subset of possible sORFs maybe indeed translated and encode a relevant peptide. sORFs located in proximity to 'regular' ORFs play and important role in regulating translation of the main ORF in many mRNA. Thus the important of sPEP products from sORFs has widely been disputed.

Comparative genomics can predict functionally relevant regions in coding and non-coding RNAs that may give rise to sPEPs by cross-species conservation. However, predictive power for such short sequences is low. Moreover, on the premise that short reading frames may be fast evolving (as are non-coding RNAs), cross-species conservation is not a strong predictor for functional importance. Nevertheless, such bioinformatic approaches can provide an upper estimate of the potential number and diversity of sPEPs in the entire human genome. Experimental approaches provide more direct evidence for translates sORFs, but the number of experimentally validates sPEPs is relatively small compared to even conservative predictions. Two methods are exquisitely useful for identifying sPEPs: Ribosome profiling identifies RNA sequences that are bound by ribosomes and thus delineates translated sORF. Direct evidence for the existence of sPEPs is provided by mass spectrometric detection of fragments or intact polypeptides (peptidomics).

The combination of the approaches above has already allowed confirming the existence of hundreds of sPEPs. The most comprehensive mass spectrometric studied conducted to date found peptide evidence for 1259 sPEP, many exhibiting cell type-specific expression. In conclusion, it is clear to date that the universe of sPEPs is large, yet unexplored.

We are looking for a highly motivated bioinformatics student with programming skills in Perl or Python, and knowledge in protein biology.

Contact

simon.elsasser@scilifelab.se
Tel 0852481227
www.elsaesserlab.org