Assignment 4 – Bioinformatics Programming Challenges – Pablo Catarecha

January 3rd, 2023.

Introduction

This assignment requires that the students perform a series of BLAST searches using Bioruby's built-in Bio::BLAST class, to retrieve reciprocal best hits between sequences of *Schizosaccharomyces pombe* and *Arabidopsis thaliana*.

Obtaining a reciprocal best hits list between two organisms is usually the starting point for an orthology search. BLASTed sequences of one organism whose best hit on other organism retrieves the original sequence when BLASTed back are considered to be orhologous, that is, sequences that share a common ancestor and that diverged with speciation events along the lineages of both organisms.

Procedure

The starting point of this assignment are two databases: one with expressed proteins from *S. pombe*, and the other with nucleotide coding sequences from *A. thaliana*. Both are provided as .zip files that have to be unzipped before executing the script.

The straightforward approach would be to perform a tblastn search using protein entries from *S. pombe* to nucleotide sequence database from *A. thaliana*, to subsequently perform the reverse blastx search using *A. thaliana* nucleotide sequences to query *S. pombe* protein database. This approach is, however, computationally resource- and time-consuming, as BLAST internally translates all nucleotide sequences into protein, in all six reading frames, and then performs a protein-protein blastp against all of them.

The approach used in this assignment seizes the concept of orf as the nucleotide coding sequence of a single protein. Within this scope, the orf database from *A. thaliana* has been translated into its protein representation, prior to performing regular single-sided blastp searches. This has led to a complete forward and back search in about two hours. If the nucleotide database had represented different features, such as contigs or any non-restricted, non-annotated sequences, this approach would not have been possible.

The databases have been imported as Bio::FastaFormat objects, taking advantage of the methods already available to manipulate the metadata relative to each entry, and using actual sequence information as a Bio::Sequence object.

BLAST factories have been built according to Bioruby documentation, adding the critical parameters that would lead to biologically relevant results[1]. Almost all critical parameters have been left as default, except e-value and output sort order. According to the bibliography, it has been demonstrated that an e-value of 10e-6 would lead to significant reciprocal best hits while maintaining an optimum level of false negatives[2]. Additionally, in the case of homology search, BLAST score is preferred to e-value when ranking the best results[3], and a bit_score above 50 would be enough to filter significantly homologous proteins of a regular size[4].

Using the above options, roughly 2300 proteins share reciprocal best hits between *S. pombe* and *A. thaliana* databases, so they can be considered orthologs in the scope of this exercise.

What next?

While the former is a generalized assumption, strictly speaking BLAST searches perform a pure similarity comparison. Orthologous genes are defined as having a common ancestor and

diverge by speciation, and BLAST cannot make a distinction between these and paralogous genes, which arose by duplication, for example.

The missing information needed to evaluate whether the reciprocal best hits found are true othologs could come from some different tests.

Adding information of additional species and building a tree with ClustalW, for instance, would allow to place BLAST hits along the evolutionary axis and help to interpret the relationship between them. Additionally, synteny analysis would reveal if the matched hits do share a neigboring genetic environment that would point to an ancestral origin.

Bibliography

1. Ladunga, I. Finding homologs in amino acid sequences using network BLAST searches. *Curr. Protoc. Bioinforma.* **Chapter 3**, Unit 3.4 (2003).

2. Ward, N. & Moreno-Hagelsieb, G. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? *PloS One* **9**, e101850 (2014).

3. Moreno-Hagelsieb, G. & Latimer, K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinforma. Oxf. Engl.* **24**, 319–324 (2008).

4. Pearson, W. R. An introduction to sequence similarity ('homology') searching. *Curr. Protoc. Bioinforma.* **Chapter 3**, Unit3.1 (2013).