Improved protein sequences for non-model organisms

Background: Molecular-genetic analyses have historically been limited to broadly studied organisms, most often those with medical or agricultural relevance. Thanks to new sequencing technologies (454, Illumina, Solid…) generating data that a few years ago would have cost millions of francs and the labor of hundreds of people over several years are now accessible to individual researchers working on non-model organisms that are interesting for their ecology or evolution.

However, a limitation of the technologies is accuracy: While less than 0.1% errors are expected in a Sanger-sequencing read, the new technologies can have error rates of between 1 and 5%.

We propose to develop a tool that can fix two types of errors found in mRNA sequence data:

- insertion/deletion errors leading to frame-shifts
- substitution errors leading to premature STOP codons

The corrections can be based on several external lines of evidence:

A. protein-level alignments with sequences from the UniProt database
B. independently obtained genomic sequences
C. identification of likely Open Reading Frames

Inputs:

a. Required: A FASTA file containing cDNA sequences.
  - *Note: sequences may contain 3' and 5'UTRs (untranslated regions), and some will contain no protein-coding regions.*

b. Optional: A BLASTX report file describing the best alignment ("hit") between each sequence from **a.** and Uniprot.

c. Optional: Genomic data for the same sequences (à discuter).

Outputs:

- A FASTA file containing all sequences (with corrections applied when necessary).
- A table with one line per correction, summarizing what was done (Each line should include: sequence identifier, evidence that a correction was needed (A,B,C; see above), location of the correction, nature of the correction (i, ii, iii; see below).

Detailed project suggestion:

1. Prepare test dataset:
    - download some mRNA sequences from NCBI.
    - Insert "errors" into some of these sequences:
        i. Substitution to premature STOP codon
        ii. Homopolymer insertion/deletion (eg: the sequence CCCCCC is read as CCCCC or CCCCCCCC – such errors are typical of sequence from 454 machines)
        iii. Randomly insertion/deletion not due to a homopolymer.
    - Prepare expected summary table.

2. Read FASTA of modified mRNAs.
3. Run BLASTX against the "nr" database (See Biopython Cookbook section "Running BLAST over the Internet"; perhaps save the result to disk to make things faster).
4. From BLASTX output, identify mRNAs that need to be corrected and where the correction should be:

    Evidence for ii and iii: a strong blastx hit ($E < 10^{-5}$) in two or more different reading frames larger than 50 amino-acids

    Evidence for i: a stop codon in the protein-level alignment more than 30 residues from the 3' end
5. Perform correction:

    If ii or iii: Scan for upstream homopolymer (repeat of more than 5 of the same nucleotide). If found, perform minimal appropriate correction.

    If i: perform correction with random nucleotide.

    Check that the correction worked!
6. Output things
7. Check that the corrected output is identical to what you expect from 1. (This should be done automatically by default)
8. Run on a larger, true dataset.

Please keep in mind that:

- no numbers or filenames should be hard-coded
- it is easiest to code if you cut things up into small functions and subfunctions that you can test individually.
- Gray cases can be considered to make the tool complete once everything else works.

Supervisor: Yannick Wurm (Keller Lab, Department of Ecology and Evolution, Biophore, bureau 3106)