Refgs.pl (ver.1.0)

Refinement of gene structure prediction by iteration

Osamu Gotoh
<<o.gotoh@aist.go.jp>>

Computational Biology Research Center (CBRC)
National Institute of Advanced Industrial Science and Technology (AIST)
AIST Tokyo Waterfront Bio-IT Research Building
2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan

Original version: June 2 2014

Overview

`Refgs.pl` is a perl script to perform refinement of eukaryotic gene structure prediction by iteratively applying `aln` for prediction and `prrn` for assessment.

Run refgs.pl

The most typical usage of refgs.pl is as follows:

```
% refgs.pl –sBefore –tAfter –I# [-R|-R1] [other options] MSA
```

`Before:` The directory that stores the predicted amino acid sequences before refinement. Each sequence should be supplemented with exon-intron structural information of the corresponding parental genes

`After:` The directory where the results of refinement will be stored.

`#:` `Maximum number of iterations`

`none:` Template selection mode M1 (minus one, code = 0)

`-R:` Template selection mode CR (closest reliable, code = 1)

`-R1:` Template selection mode PR (profile of reliable, code = 2)

`-a`*N*: Significance level of outlier analysis (default = 0.1)

`-k:` In the default setting, the sequences judged to be pseudogene products are eliminated from updated MSAs. With the –k option, all sequences in the input are kept unremoved in the updated MSAs.

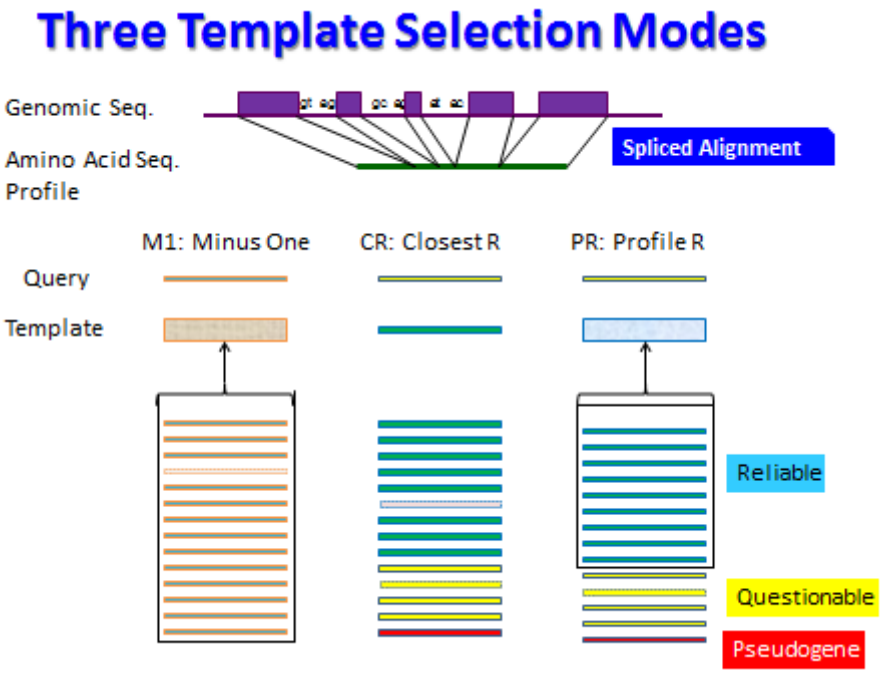
`-l`*N*: The length of the left margin from the current translation initiation site to be analyzed by the next cycle of `aln`.

`-r`*N*: The length of the right margin from the current translation termination site to be analyzed by the next cycle of `aln`.

`-uA:` Diagnose reliability of each predicted gene structure.

`MSA:` Gene-structure-aware multiple protein sequence alignment (GSA-MPAS) constructed from aa sequences in `Before` directory

At each refinement step, new MSA files, `MSA.1`, `MSA.2`, … will be generated. The last file `MSA.`*N* is constructed from the updated amino acid sequences in `After`. The three template selection modes are illustrated in the next page.



System

To run `refgs.pl`, `aln`, `prrn`, and several other programs, as well as several perl scripts, must be properly installed. Please refer to the documents of `aln` or `prrn` for install of these programs. Moreover, you must prepare `seqdb` directory where indexed and formatted genomic sequences are stored and `table` directory where various tables used by `aln`/`prrn` are stored. `make install` command will generate these directories under your home directory by default, but you can change the locations of these directories at the time of `.configure`. It is also possible to specify the locations of `seqdb` and `table` by `env` variables `ALN_DBS` and `ALN_TAB`, respectively.

Sequence files

Genomic sequences

The genomic sequence should be provided in the generally used FASTA format. However, `refgs.pl` imposes a few specific rules.

1. The file name should be *xxxxxxxx*_g.gf, where *xxxxxxxx* is an eight-character identifier of the genome. Each *x* should be in lower case.
2. Each entry within a FASTA file should have the form >*Xxxxxxx*…. The first eight characters must be the same as the first eight characters of the file name, except that the first character should be upper case. The part … may be composed of any non-space characters, but they must be unique within each file. There is no length limitation in the entry ID, but excessively long ID would be inconvenient in later processes.
3. `% makeidx.pl` *xxxxxxxx*`_g.gf`
   command in the `seqdb` directory will generates several files necessary for access to

the genomic sequences by `aln`.

Gene-structure-supplemented amino acid sequences

The amino acid sequences given in the `Before` directory should contain information about the gene organization of the parental gene. The extended FASTA format described in the documents of `aln` or `prrn` describe the details of the format.

GSA-MPSA

When Seq1, Seq2, …, Seqn are gene-structure-supplemented amino acid sequences included in the `Before` directory, the command

```
% prrn -s Before –o MSA Seq1  Seq2 … Seqn
```

will generate the initial GSA-MPSA to be refined.

`Refgs3.pl`

As described in Ref. 1, a combination of different template selection modes improves overall quality of refinement by `refgs.pl`. `Refgs3.pl` is designed to facilitate this procedure. In `Refgs3.pl`, the directory names are fixed; `prd`, `prd1`, `prd2`, `prd3` are used as `Before` and consecutively refined `After` directories. The initial MSA before refinement should be named `MSA.`*n*, where *n* is a numeral. The final MSA file is named `TRD.`*n*. All intermediate MSA files are reserved. Alternatively, you may give a UPGMA tree file as the argument. In that case, the initial MSA will be generated from the sequences in the `prd` directory using the given tree as the guide tree.

```
% refgs3.pl [-sBefore] [-mfst] upgma_tree.n
% refgs3.pl [–f1] [-sBefore] [-mfst] MSA.n
```

By default, a series of CR-M1-PR (102) template selection modes are used. The `–mfst` option can change this series, where *f*, *s*, and *t* are codes (0, 1, or 2) shown above. If only `f` or `f` and `s` are specified, the series will be stopped after the first or the second stages.

When the initial MSA is given or constructed, `refgs3.pl` tries to remove redundant sequences and minor isoforms derived from the same genomic region. With –f1 option, this step can be skipped.

References

[1] Gotoh, O., Morita, M. Nelson, D.R. (2014) "Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment", submitted

---

Osamu Gotoh <o.gotoh@aist.go.jp>

Last modified on 2014-06-02