alfie

A package for nuclear anonymous loci prediction and phylogenomic analysis using complete genomes!

Thank you for downloading Alfie!

SUMMARY

1. Introduction
    1. What is alfie?
    2. How it works?
    3. What to use alfie for?
2. Installation
    1. Downloading
    2. Installation requirements
    3. Testing your instalation
3. Quick Run
4. Advanced Parameters
5. Citing Us

1. Introduction

1.1. What is alfie?

Alfie is an open-source python package containing a pipeline thatsearchs for single-copy, neutral evolving and non-linked anonymous loci.

In other words, it looks for non-coding loci with a set of characteristics which makes them perfect for phylogenomics and populational genetic analyses. The goal of alfie is to be a complete, user-friendly and flexible anonymous loci predictor.

1.2. How it works?

Alfie utilises a sequential modular approach, with each step of the pipeline being stored in a single module that can be excecuted with greater flexibility in standalone mode. Many of these steps require external programs, such as BLAST or CLUSTALW. To reduce the complexity of the steps the whole pipeline can be run automatically using the afie.py script.

Using as input a single query genome and its associated GTF file, alfie finds all anonymous regions distant from genomic features (>200kb distance, by default) to avoid the effect of genetic linkage. After the initial anonymous regions have been found, they are split in small pieces (default 1kb). Each piece is searched against subject and query genomes to filter duplications and find orthologous regions in all genomes.

Alfie performs an exhaustive search, i. e., it finds every possible anonymous regions for the specified parameters in the genomes analyzed. Using default parameters and closelly related vertebrate genomes, alfie will usually find 100-1000 anonymous loci. Each ortholog group of anonymous region is written in NEXUS, PHYLIP, ALN and FASTA formats. Alfie will also concatenatenate all the alignments into a large file for phylogenomic analysis. The package comes with support for running phylogenomics and population genetics software.

1.3. What to use alfie for?

Alfie was developed to find hundreds to thousands of independent, nuclear markers among whole complete genomes and facilitate population genomics studies. An usual user case would use four to ten closely related genomes (we recommend no more than 20 million years of divergence to allow precise ortholog assignment). As an example, we were able to find about 300 anonymous loci with ortologues in all four hominoid genomes used (human, chimpanzee, gorilla and orangutan).

It is strongly recommended that at least one of the genomes have been extensively studied and annotated, presenting a comprehensive GTF file that describes the precise location of the main features targeted by natural selection, such as genes and regulatory elements.

2. Installation

2.1. Downloading alfie

Users can either download and install the entire source code from github at the address https://github.com/igorrcosta/alfie/archive/master.zip or download the docker container with all the dependencies included (available soon). To install, simply extract the donwloaded repository file in a folder.

2.2. Installation requirements

Working versions for each required program can be found on the following links:

- Blast: http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=Blast Docs&DOC_TYPE=Download
- ClustalW: http://sourceforge.net/projects/mira-assembler/
- PhyML: http://www.atgc-montpellier.fr/phyml/binaries.php
- ModelTest: https://code.google.com/p/jmodeltest2/
- Biopython: http://biopython.org/wiki/Download

2.3. Testing your installation

You can test your installation by running alfie on the human and chimpanzee Y chromossomes' test case. To do that, go to the alfie folder and run the following command:

```
$>python alfie.py -i test/homo_y.fasta test/pan_y.fasta -g test/y.gtf
```

This command might take more than a minute to finish, depending on your computer speed. A successful run will output 42 candidate loci in the test.fasta file, of which only loci 11, 27 and 40 will be selected and aligned.

3. Quick Run

To run alfie, you will need a reference genome and gtf file (there are several available at http://www.ensembl.org/info/data/ftp/index.html) and some genomes to compare against the reference. Example command line:

```
$> python alfie.py -i "reference_genome.fasta" "genome2.fasta" "genome3.fasta" -g
"annotation_file.gtf" -o "output_folder"
```

Here the ".fasta" represents the path to the genome files in FASTA format, "annotation_file.gtf" is the path to the gtf file and "output_folder" is where the loci will be saved.

This command will find several candidade loci in the reference genome, which will be stored at a test.fasta file. This candidade loci will be blasted against all genomes and the final loci will be saved in several formats.

4. Advanced parameters

While alfie can be executed with good results without any additional configuration, there are several options to flexibilize your analysis:

Flag	Description
-i, --genomes	Path to the genome files in the FASTA format. The first genome file inserted will be considered the reference genome.
-g, --gtf	Path to the gtf file relative to the reference genome.
-o, --outpath	Path where all output files will be saved.
-f, --skip_formatdb	Skip making BLAST databases, use databses fromthe lastrun. Can significantly improve analysis speed.
--locus_length	Length of the anonymous loci (default 1000bp). You may increase this length if you are planning to predict primers for these loci.
--max_n	Maximum percentage of N's in the AL sequence (default 0%). Increase this if you want to find more loci in a low quality reference genome.
--inter_distance	Minimum distance between ALs (defaut 200000bp).
--gene_distance	Minimum distance between ALs and genes (default 200000bp). You can use negative numbers to find loci close to the gene regions, for example, -2000 will find loci between 0-2000bp from genes
--end_distance	Minimum distance between ALs and the telomeres (default 10000bp).
--gene_locus	Use this flag to find loci inside the gene regions (will ignore the gene_distance flag).
--cds	Only considers the CDS features of GTF files, ignoring all pseudogens, miRNA, etc.
--duplication_cutoff	ALs with 2 hits with identity higher than this will be considered duplicated(default: 50%).
--identity_cutoff	ALs with a identity higher than this will be considered homologous(default: 90%).
--coverage_cutoff	BLAST hits must have at least this much %coverage to be considered hits (default: 90%).

Flag	Description
--chromossomes	Chromossomes to be excluded. We recommend excluding all sex chromossomes.
--min_align	Minimum final alignment length (default 900bp). This will exclude loci that have many Ns in genomes other than the reference.
--remove_gaps	Remove gaps from the final alignment.

5. Citing us

If you use this program on your analysis, please cite the following paper:

Costa IR, Prosdocimi F, Jennings WB (2016). In silico phylogenomics using complete genomes: a case study on the evolution of hominoids. Manuscript in preparation.