OMA stand-alone

CBRG, ETHZ

Contents

1 Introduction

You can download and install OMA as a stand-alone version. Included are the algorithms for OMA itself plus its addition ESPRIT. The software can be installed on Linux (x86, both 64bit and 32bit) and MacOSX (x86, both 32bit and 64bit).
For more information about OMA and ESPRIT in general, please have a look at the OMA browser page:
http://omabrowser.org/Algorithm.html
If you have specific questions about the installation or the usage of OMA, please contact {adrian or cdessimoz}@inf.ethz.ch .

2 Downloads

The current version of OMA stand-alone can be found here:
OMA.0.99m.tgz

3 Installation

You do not need to install OMA stand-alone on your system; the script will also run if you just call it by using the complete path to `bin/oma` in the installer folder. But we still encourage you to run the installer script, since it makes working with OMA considerably more convenient.
To install OMA stand-alone on your system, download the installer, untar the package and run the included installer script:

```
curl http://omabrowser.org/standalone/OMA.0.99m.tgz -o oma.tgz
tar xvzf oma.tgz
cd OMA.0.99m
./install.sh /your/install/prefix
```

If you do not choose an install prefix, OMA will be installed in `/usr/local/OMA` (for this, you might need to install it using the root account or `sudo`).
After installation, make sure the `bin` folder of OMA is in your `PATH` variable, e.g., if you are using bash and used `/your/install/prefix` as installer prefix, add a line in `/.profile` such as:

```
export PATH=\$PATH:/your/install/prefix/OMA/bin
```

For other shells, choose the appropriate syntax.

4 Usage

First, set up a working directory. Copy the file `parameters.drw` into this folder and change it to your needs. Create a directory `DB` in your working directory that holds the genome data in FASTA format (see 'File formats') and copy your data into this directory. If you want to use ESPRIT, the FASTA file containing the contigs should be called `{YourGenome}.contig.fa`.
Then, simply call `OMA` from your working directory to run OMA and/or ESPRIT
If you have not installed OMA yet, use the complete path to `bin/oma` in the installer folder to start the script.
As an example, assume you installed `OMA` in `/your/install/prefix` and want to use ESPRIT on two genome files and one file with contigs (all in `/home/you/fasta`, do something like this:

```
# create working directory
mkdir myWorkingDir
cd myWorkingDir
```

```
# create DB directory in working directory
mkdir DB
# copy FASTA files into DB directory
cp /home/you/fasta/yourFirstGenomeFile.fa DB/
cp /home/you/fasta/yourSecondGenomeFile.fa DB/
cp /home/you/fasta/yourContigFile.contig.fa DB/
cp /your/install/prefix/OMA/OMA.0.99m/parameters.drw ./
# adjust parameters
vim parameters.drw
# run OMA
OMA
```

To get a first impression of `OMA` you could `cd` into the `ToyExample` directory, have a look at `parameters.drw` and run `OMA` to process our example files.

4.1 Parallelization

The all-against-all phase of OMA is the most time-consuming one, but it can be parallelized (unlike all other steps, which cannot run in parallel). The way it works is that the parameter "NP" and the total number of genomes (n) will determine into how many chunks the all-against-all phase is divided. If NP=1, there will be n*(n-1)/2 parts, if NP is higher, there will be more (shorter) jobs.
Now, scheduling is straightforward: all compute processes need to start from the same directory, and will try to do all the chunks sequentially. However, before starting a new chunk, each process ensures that it has not yet been claimed/processed by another process (i.e. no result file yet exists).
Therefore, there is not need to specify which parts are to be done by which process. One should only ensure that all processes start from a shared directory, such that each chunk gets computed by a single process only.

5 File Formats

5.1 Input Files

OMA uses two different input formats: FASTA files for genome input and a Darwin file for parameter input.
The Fasta format is explained in detail on wikipedia.
OMA uses the greater-than symbol '>' to distinguish labels from sequences (in contrast to the possibility of using a semicolon ';'). Each sequence in an MSA is supposed to have its own label. Have a look at the FASTA files included in ToyExample/DB in our installer package for some example files.
If you want to use ESPRIT, make sure that FASTA files containing contigs are called `{YourGenome}.contig.fa`. So if you want to experiment with some mouse genome, call the FASTA file `mouse.contig.fa` or `mymouse.contig.fa` or something similar.

Parameter files use Darwin syntax. Key-value-pairs are written as

```
key := value;
```

Note the colon in := and the semicolon at the end of the line. If your parameter file does not use valid Darwin syntax, OMA will print out a short message and stop its execution.

5.2 Output Files

5.2.1 OMA Output

The output of OMA gets written to files stored in a folder `Output` in your working directory. There are three text files plus an additional folder `PairwiseOrthologs` that contains one file for each pair of your genome sets.
The textfiles are organized as described in Table 1.

Filename	Contents
`Map-SeqNum-ID.txt`	Lists all genes of all datasets with their unique sequence number and the labels read from the FASTA files.
`OrthologousGroups.txt`	The groups of orthologs are given as one per row, starting with a unique group identifier, followed by all group members, all separated by tabs.
`OrthologousMatrix.txt`	More compact version of `OrthologousGroups.txt`. The groups of orthologs are given as matrix with group per row and one genome per tab-separated column. Numbers refer to entry number as listed in the file `Map-SeqNum-ID.txt`.

Table 1: Contents of the OMA output files

The textfiles in `Output/PairwiseOrthologs` are named according to {genome a}-{genome b}`.txt` and consist of a list of pairwise orthologs for the two given genomes. Every pair is listed only once, and in no particular order. Each line in the file contains one pair; all fields are separated by tabs. In the first two field, the unique IDs of the proteins are given. The next two fields contain the labels of the proteins, and in the last two fields, the type of orthology and (if any) the OMA group is given.

5.2.2 ESPRIT Output

ESPRIT stores its output files in a folder`EspritOutput` in your working directory. The output consists of three text files and one tarball. In the tarball, FASTA files with the MSAs of the hits ESPRIT found are stored. The other three files are explained in detail in Table 2.

Filename	Contents
`params.txt`	This file is kept as a reference and contains all parameters used in the current run.
`hits.txt`	All hits found by ESPRIT are listed in this file. It is a list of contigs, ordered according to their position relative to the putative ortholog. Each line describes one contig, the fields are separated by tabs. In the first field, the fragment pair ID is printed; the next two fields contain the labels of the first and second fragments found in this hit. The forth and fifth fields contain the label of the corresponding full gene and its genome name. Then follows the distance difference between the two fragments and the number of positions between them (i.e. the gap); at last, an array is listed containing the IDs of all s3 genes corresponding to this hit.
`dubious.txt`	ESPRIT often detects more candidate pairs than it will list in the `hits.txt` file, but not all of them survive the quality check. Still, if you want to see which triplets have been filtered out, have a look at `dubious.txt` where they are still listed. The file format is the same as for `hits.txt`.

Table 2: Contents of the ESPRIT output files

6 Parameters

All parameters for OMA and/or ESPRIT are set in a parameters file. There is an example file in the OMA installer package; we encourage you to copy this file into your working directory and change it to your needs.
The parameter file consists of two main parts: First, general parameters for OMA are set; see Table 3 for detailed explanations. Second, more specific parameters that only affect the ESPRIT algorithm can be changed. These parameters are explained in Table 4. Note that changing the ESPRIT parameters will not have an effect unless you set the boolean variable `UseEsprit` to `true`.

7 License

Parameter	Meaning	Default
ReuseCachedResults	If you want to recompute everything from scratch every-time the script is run, set this to false.	true
NP	In the all-against-all phase, each genome pair is split in NPparts. This allows for shorter milestone steps (e.g. in case of computer crash) and allows to parallelize jobs with few but large genomes.	4
MinScore	Alignments which have a score lower than MinScore will not be considered. The scores are in Gonnet PAM matrices units.	181
LengthTol	Length tolerance ratio. If the length of the effective alignment is less than LengthTol * min(length(s1), length(s2)), then the alignment is not considered.	0.61
StablePairTol	During the stable pair formation, if a pair has a distance provable higher than another pair (i.e. StablePairTol standard deviations away) then it is discarded.	1.81
VerifiedPairTol	Length tolerance ratio. If the length of the effective alignment is less than LengthTol * min(length(s1), length(s2)), then the alignment is not considered.	1.53
MinSeqLen	Any sequence which is less than MinSeqLen amino acids long in regular genomes is not considered.	50

Table 3: General parameters in OMA

Parameter	Meaning	Default
UseEsprit	You can either set this to true, which will enable ESPRIT and shut down the parts of OMA that are not directly needed for ESPRIT, or set it to false to make no use of ESPRIT at all.	false
DistConfLevel	Confidence level variable for contigs. This is the parameter tol described in the paper.	2
MinProbContig	Minimal proportion of genomes with which contigs form many:1 BestMatches to consider that we might be dealing with fragments of the same gene. This is the parameter MinRefGenomes described in the paper, normalized by the total number of reference genomes.	0.4
MaxContigOverlap	Maximum overlap between fragments of same gene from different contigs.	5
MinSeqLenContig	Any sequence which is less than MinSeqLenContig amino acids long in contigs is not considered.	20
MinBestScore	Minimum best score for BestMatch in scaffold recognition.	250

Table 4: ESPRIT parameters