SABmark documentation
Version: 1.65

Ivo Van Walle*

September 8, 2004

*ivwalle@vub.ac.be, Vrije Universiteit Brussel

Contents

1 Introduction

The Sequence Alignment Benchmark (SABmark) contains alignments that cover the entire known fold space, as classified by SCOP.[2] To limit the impact of highly abundant folds, each alignment contains at most 25 sequences. Currently 2 alignment sets are available, Twilight Zone and Superfamilies, which represent sequences with resp. very low to low, and low to intermediate similarity. These are based on subsets provided by the ASTRAL compendium and correspond roughly with 0-25% and 0-50% identical residues.[1] Alignments of sequences with higher similarities are not provided, since the performance of most algorithms if already very good above 50% identities.

Since many alignments are performed exactly to determine whether or not sequences are related, a second version of both the Twilight Zone and the Superfamilies set is given that addresses this issue: to each group of sequences to be aligned, the same number of 'false positive' sequences (sequences that belong to a different fold) is added.

2 Database structure

The directory structure of SABmark is organised as follows:

```
SABmark/ (database home directory)
    scripts/
        archive.pl          : archive data of some set/run
        archive_all.pl      : archive all unarchived data of some set
        extract.pl          : extract data of some set/run
        remove.pl           : remove all data of some set/run
        remove_all.pl       : remove all data of all runs of some set
        score.pl            : get scores for some set/run
        score_all.pl        : score all unscored runs of some set
        score_alignments*   : used by score.pl
        bl2seq_to_alignm.pl : converts Blast 2 sequence report to Alignm format
        run_*.pl            : scripts to launch each run
        dblist.pl           : overview of all the sets/runs
    twi/ (set directory)
        archive/  : .tar.gz file for each run
        analysis/ : scores for each run
        group1/ (group directory)
            fasta/            : separate Fasta files for each sequence
            reference/        : pairwise Fasta alignment files
```

```
        pdb/              : PDB files for each sequence
        "run" directories : files generated by some program
        group.fasta       : all sequences of the group
        group.summary     : some data about each sequence
    ...
    groupN/
    set.fasta   : all sequences of the set
    set.summary : some data about each group
sup/
twi_fp/
sup_fp/
db.fasta    : all sequences of the database
db.summary  : some data about each set
SABmark.pdf : documentation
```

3 Typical use

After installing SABmark, the typical tasks needed to analyse some program / parameter settings (together a "run") are:

1. Make a script file that will execute the run for each group of a set. The template scripts `run_alignm.pl.template` and `run_bl2seq.pl.template` can be used as a starting point.
2. Execute the script on some or all sets.
3. Get scores for all the created alignments with `score.pl` or `score_all.pl`.
4. Archive all data with `archive.pl` or `archive_all.pl`.
5. Further analyse the data in $setdir/analysis/$runname.scores, and/or generate additional scores.

4 Scoring alignments

All the alignments of a given run on a given set are scored by a singe command:

```
score.pl $set $run
```

This will create a file `$setdir/analysis/$run.scores` and a file `$setdir/analysis/$run.times`. For the script to work, the output alignments in each group's directory must have the `.alignm` (see [4]), `.fasta` or `.clustal` extension and of course also have the corresponding formats. The scores file contains for each sequence pair a single line with tab delimited data. The fields are as follows:

1. Group id
2. Flag set to 1 if the first sequence is a true positive
3. Flag set to 1 if the second sequence is a true positive
4. Length of the first sequence
5. Length of the second sequence
6. Structure similarity as defined by SCOP (Sl = Scop level): 0 (different class), 1 (different fold), 2 (same fold only), 3 (same superfamily only), 4 (same family only), 5 (same domain)
7. Name of the first sequence

8. Name of the second sequence
9. Length of the reference alignment
10. Length of the test alignment
11. Number of residues aligned correctly in the test alignment
12. Percentage identity of the reference alignment
13. f_D score
14. f_M score

The `score.pl` script calls the program `score_alignments`, which compares pairwise test alignments to pairwise reference alignments, and which may both be given also as a multiple alignment. It can be used separately as well, and the command line arguments are given in table 1. The output is printed to STDOUT and consists of the same fields as in the scores file (see list above), starting from the name of the first sequence.

Table 1: Command line arguments for score_alignments

Parameter	Value
`-r`*	Reference alignment file
`-t`*	Test alignment file
`-s`*	Sequence file (any alignments in it are not used)
`-optfile`*	File with command line arguments (e.g. in case the command line would become too long otherwise)
`-header`	No value. A header line with the name of each value will be printed
`-pair_subset`	Determines which sequence pairs should be scored. 'ALL': all pairs for which a reference or a test alignment is given. 'REF': all pairs for which a reference alignment is given. 'TEST': all pairs for which a test alignment is given. 'COMMON': all pairs for which a reference and a test alignment are given
`-copyright`	A copyright message will be printed
`-separate_alignm_records`	Applies to Alignm format files: in case multiple records contain the same sequence pair, each will be considered as a separate alignment

*More than one may be given.

5 Installation

1. Extract the tarball to the database directory of choice
2. Extract the reference alignments of each set with `$dbdir/scripts/$extract.pl $set reference`.
3. When needed, do the same for the individual Fasta files (archive name: `fasta`).
4. PDB files for all sequences are provided separately because of their size. After downloading, copy the tarball to the SABmark home directory and unpack it. A file `pdbs_install.script` and a directory `pdbs_install` will be extracted. Then run `source pdbs_install.script` and remove the file and directory after. The PDBs wil now be present in each group directory and also as an archive.

References

[1] John-Marc Chandonia, Gary Hon, Nigel S Walker, Loredana Lo Conte, Patrice Koehl, Michael Levitt, and Steven E Brenner. The ASTRAL Compendium in 2004. *Nucleic Acids Res*, 32 Database issue:D189–92, Jan 2004.

[2] AG Murzin, SE Brenner, T Hubbard, and C Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, Apr 1995.

[3] JM Sauder, JW Arthur, and RL Dunbrack. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, 40(1):6–22, Jul 2000.

[4] Ivo Van Walle, Ignace Lasters, and Lode Wyns. Align-m–a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics*, 20(9):1428–35, Jun 2004.