HOMOLOBIND users guide v1.1

Fred P. Davis
Janelia Farm Research Campus
Howard Hughes Medical Institute
davisf@janelia.hhmi.org
http://pibase.janelia.org/homolobind

November 24, 2010

Contents

1 Introduction

HOMOLOBIND identifies residues in protein sequences with significant similarity to structurally characterized binding sites. The program predicts residues in ligand and protein binding sites with estimated true positive rates of 98% and 88%, respectively, at 1% false positive rates.

HOMOLOBIND transfers binding sites from LIGBASE (`http://salilab.org/ligbase`) and PIBASE (`http://pibase.janelia.org`) through ASTRAL/ASTEROIDS (`http://astral.berkeley.edu`) alignments onto SUPERFAMILY (`http://supfam.org`) domain assignments.

This release is based on the SCOP v1.73 domain classification. SUPERFAMILY has recently (Nov 2010) updated to SCOP v1.75 domain definitions, and the corresponding HOMOLOBIND update will be available by the end of 2010.

2 Installation

2.1 Download HOMOLOBIND

The package is freely available at `http://pibase.janelia.org/homolobind`. The main software file is homolobind_v1.1.tar.gz. The server also provides data files that the program downloads automatically.

2.2 Prerequisites

1. Perl

   HOMOLOBIND requires Perl (`http://www.perl.org`), usually installed by default on Mac OS X and GNU/Linux machines. Windows users can install Perl using the Cygwin package (`http://www.cygwin.com`).

2. Bit::Vector perl module from CPAN: `http://search.cpan.org/search?query=bit-vector&mode=all`

3. System utilities: wget, tar, gunzip

4. Optional: R (`http://r-project.org`) and the R perl interface (`http://www.omegahat.org/RSPerl/RFromPerl.html`) are required to calculate the significance of overlap between predicted ligand and protein binding sites.

2.3 Installing HOMOLOBIND

Get the package and uncompress it in your favorite location:
`http://research.janelia.org/davis/homolobind/files/homolobind_v1.1.tar.gz`

```
mv homolobind_v1.1.tar.gz /your/favorite/location
cd /your/favorite/location/
tar xvfz homolobind_v1.1.tar.gz
```

1. Place the src/perl_api directory in your PERL5LIB environment variable.

   For example, if you run a csh or tcsh shell, add this to your .cshrc file:

```
setenv PERL5LIB
   {$PERL5LIB}:/your/favorite/location/homolobind_v1.1/src/perl_api
```

   For a bash shell, add this to your .bashrc:

```
PERL5LIB=/your/favorite/location/homolobind_v1.1/src/perl_api
export PERL5LIB
```

2. Edit the src/perl_api/homolobind.pm specs section (line 70) to set the directory where data files should be stored

3. If you want to run the program in parallel on an SGE-based computing cluster, edit the homolobind.pm specs section to:

   - specify the hostname from where jobs can be submitted (line 64)
   - the number of jobs to launch (line 65)
   - the qstat polling frequency (line 66)

```
   $homolobind_specs->{SGE}->{headnode} = 'login-eddy' ;
   $homolobind_specs->{SGE}->{numjobs} = 25 ;
   $homolobind_specs->{SGE}->{qstat_sleep} = 120 ;
```

4. Run HOMOLOBIND to retrieve the required data files (~520MB).

```
perl homolobind.pl -fetch_data 1
```

5. Download SUPERFAMILY self-hits files:

   (a) Retrieve the tar file:
   http://pibase.janelia.org/download/homolobind/v1.1/self_hits_1.73.tar.gz

   (b) Uncompress the file in the 'superfamily' subdirectory of the data directory specified in Step 2.

```
mv self_hits.tar.gz HOMOLOBIND_DATA_DIRECTORY/superfamily/
cd HOMOLOBIND_DATA_DIRECTORY/superfamily/
tar xvfz self_hits.tar.gz
```

3 Running HOMOLOBIND

Note, examples/README.examples describes how to recreate Fig. 4 and 5 in the accompanying manuscript. The directory includes the input file and expected output files.

3.1 Preparing input for HOMOLOBIND

HOMOLOBIND takes as input a list of SUPERFAMILY domain assignments (File format described in 'Input file formats' section). There are 2 ways to get these assignments:

1. Run SUPERFAMILY software (v1.73) locally to assign domains to your sequences: `http://supfam.org/SUPERFAMILY/downloads.html`

   NOTE: This option currently only works with v1.73 SUPERFAMILY software. SUPERFAMILY has recently (Nov 2010) updated to SCOP v1.75 domain definitions, and the corresponding update for the HOMOLOBIND binding site library will be available by the end of 2010.

2. Get precomputed genomic domain assignments by installing SUPERFAMILY MySQL tables and querying for your species of interest.

   Three tables are required – align, ass, and family – and can be downloaded here:

   - align_01-Nov-2009.sql.gz (4 GB)
   - ass_01-Nov-2009.sql.gz (318 MB)
   - family_01-Nov-2009.sql.gz (167 MB)

   After installing them into a local MySQL database, all human domain assignments can be retrieved with the following MySQL command:

```
mysql> SELECT ass.genome, ass.seqid, ass.model, ass.region,
   ass.evalue, align.alignment, family.evalue, family.px,
   family.fa FROM ass, align, family WHERE ass.genome = 'hs' AND
    ass.auto = align.auto AND ass.auto = family.auto ORDER BY
   ass.model, family.fa ;
```

3.2 Predict binding sites for a set of proteins

```
perl homolobind.pl -ass_fn ASSIGNMENT_FILE -out_fn OUTPUT_FILE
   -err_fn ERROR_FILE
```

- -ass_fn SUPERFAMILY_ASSIGNMENT_FILENAME
- -cluster_fl <0|1> - optional; if 1 will run the query in parallel using an SGE computing cluster, as specified in `src/perl_api/homolobind.pm`
- -out_fn OUTPUT_FILENAME - optional; default to STDOUT
- -err_fn ERROR_FILENAME - optional; default to STDERR

3.3 Create diagrams of predicted binding sites

```
perl homolobind.pl -plot_annotations 1 -ass_fn ASSIGNMENT_FILE
   -results_fn HOMOLOBIND_RESULT_FILE -seq_id SEQUENCE_IDENIFIER
   -seq_id_fn SEQUENCE_ID_LIST_FILE
```

This command will create postscript diagrams depiciting the annotated binding sites for the sequence specified by `-seq_id XX` or the sequences listed in the file specified by `-seq_id_fn`.

3.4 Summarize prediction results

```
perl homolobind.pl -summarize_results HOMOLOBIND_RESULTS_FILE >
   summary.txt
```

This command summarizes the annotation results: numbers of proteins, domains, and residues covered by domain, peptide, and ligand binding sites. This command also counts predicted bi-functional residues, with similarity to both ligand and protein binding sites.

- -withR <0|1> - optional; if 1, calls R through the RSPerl interface to calculate the significance of overlap between predicted ligand and protein binding sites.

4 HOMOLOBIND output file formats

4.1 Binding site predictions

Each output line describes the similarity of a target domain to a structurally characterized binding site. The output is tab-delimited:

1. Sequence identifier
2. Domain residue range
3. SCOP classification level (superfamily or family)
4. SCOP classification
5. Template binding site type: p=peptide, P=protein domain, L=ligand; exp=exposed residues.
6. Template binding site identifier.
7. Template binding site description
8. List of residues aligned to template binding site
9. Percent sequence identity over template binding site residues
10. Percent sequence similarity over template binding site residues;
    Similarity is a Karlin-Brocchieri normalized and rescaled BLOSUM62 score, equation below.

11. Number of identical binding site positions
12. Number of aligned binding site positions
13. Number of residues in template binding site
14. Fraction of template binding site residues aligned to the target sequence
15. Number of identical residues across whole domain
16. Length of whole domain alignment
17. Percent sequence identity across whole domain
18. Percent sequence similarity across whole domain

For those target domains where similar binding sites were not found, there will be a line with similar fields as above, however in place of fields 6-18, one of the following reasons will be provided:

1. 'no template'
2. 'sub-threshold template' - template binding sites are available in the SCOP family, but are below the sequence identity threshold.
3. 'domain family not covered by ASTRAL'. ASTRAL/ASTEROIDS only covers domains in classes a-g. Classes h-k are not covered.
4. 'ERROR in merging SUPFAM/ASTRAL alignments'.

To provide a more graded measure of template–target similarity, a sequence similarity score is computed using a normalized version of the BLOSUM62 substitution matrix (Henikoff and Hennikof, *Proc Natl Acad Sci*. 1992) as suggested by Karlin and Brocchieri (*J Bacteriol*. 1996) and rescaled to range from zero to one:

$$sim(aa_i, aa_j) = \left(\frac{BLOSUM62(aa_i, aa_j)}{\sqrt{BLOSUM62(aa_i, aa_i) \cdot BLOSUM62(aa_j, aa_j)}} + 1\right)/2 \quad (1)$$

4.2 Prediction summary

The output has 3 parts:

- Annotation summary for each protein (tab-delimited)
    1. 'OVERLAP_SUMMARY'
    2. sequence identifier
    3. number of exposed residues
    4. number of ligand-binding residues
    5. number of protein-binding residues
    6. number of bi-functional positions

7. Significance of ligand/protein binding site overlap (right-sided p-value); blank if -withR option not provided
8. Significance of ligand/protein binding site non-overlap (left-sided p-value); blank if -withR option not provided
9. overlap score
10. ';' delimited list of unique SCOP families in the protein, as annotated by SUPERFAMILY.

- Overall counts for number of domains/proteins/residues/families.
- Summary table in LaTeX format - similar information as above.

5 Input file formats

5.1 SUPERFAMILY domain assignments

Tab-delimited file:

1. species identifier
2. target sequence id
3. SUPERFAMILY model_id
4. target domain residue range
5. superfamily-level domain assignment e-value
6. SUPERFAMILY alignment string
7. family-level domain assignment e-value
8. SCOP px_id
9. SCOP fa_id

5.2 Substitution matrix

matblas format, eg: `http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt`

6 Citing HOMOLOBIND

Proteome-wide prediction of overlapping small molecule and protein binding sites using structure.
Davis FP. *Molecular BioSystems*, 2011. Advance Article. `http://dx.doi.org/10.1039/C0MB00200C`

HOMOLOBIND uses data from the following sources:

- ASTRAL/ASTEROIDS: Chandonia, et al. Nucleic Acids Res (2004) 32:D189-92.
- LIGBASE: Stuart, et al. Bioinformatics (2002) 8(1):200-1.

- PIBASE: Davis and Sali. Bioinformatics (2005) 21(9):1901-7.
- SCOP: Murzin, et al. J Mol Biol (1995) 247(4):536-40.
- SUPERFAMILY: Wilson, et al. Nucleic Acids Res (2009) 37:D380-6.

7 Contact Information

Fred P. Davis
Howard Hughes Medical Institute
Janelia Farm Research Campus
19700 Helix Dr
Ashburn, VA 20147, USA
email: davisf@janelia.hhmi.org
phone: (571)-209-4000 x3037