Installation and User Manual for ILbind

Before using the ILbind, a user should generate predictions from FINDSITE and SMAP methods; we include example that shows how to do that. The prediction requires: structures of target protein(s), structure of the considered ligand, and at least one template of the ligand-protein complex. The predictions must be generated on a computer with Linux OS by following these five steps:

1. Download and install FINDSITE (v. 1.0) package according to instructions available at http://cssb.biology.gatech.edu/skolnick/files/FINDSITE

2. Download and install SMAP (v. 2.0) package according to instructions available at http://funsite.sdsc.edu/scb/smap/smap.html

3. Run FINDSITE for the selected ligand using the target protein(s) and the chosen template(s) as FINDSITE's threading library (see example below)

4. Run SMAP for the selected ligand using the target protein(s) separately for each chosen template (see example below)

5. Run ILbind (see example below)

We recommend that the user prepares the target protein(s) by removing all non-amino acid molecules from the PDB file(s), and the template(s) by removing all atoms that do not belong to the protein chain and the bound selected ligand.

SMAP by default is downloading all missing pdb files into PDB folder defined in its configuration file. User has to copy the prepared files to that folder prior to running SMAP, or SMAP will use PDB version of the files. The files in PDB directory must be in lowercase and without chain identifier (i.e., 1cqi is a correct name, while 1CQI or 1CQI_B are not). Name of the template file cannot have the same name as any of the target proteins; otherwise it will be overwritten by this target protein.

ILBind uses LIBSVM, an SVM library that is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Precompiled files (for 64bit Linux distribution) **svm_predict** and **svm_scale**, are added to the ILbind distribution. In case that the files do not run on user's machine, a suitable LIBSVM distribution can be downloaded at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Example

Three templates are used for the inverse binding prediction for the ADP ligand: tmp1 (2zpa_A), tmp2 (1gzf_C), and tmp3 (1cqi_B). Our example considers ten target proteins:

- positive (ADP-binding) targets: 1cqi_b, 1c30_c, 1ce8_c, 2c31_b, 2hmf_c
- negative (not binding to ADP) targets: 1h2c_a, 1wpa_a, 1qdd_a, 1g3p_a, 2grr_b

All files for this example are included with the distribution. The targets are stored in the input folder and the templates can be found in the templates folder. The output files that are generated by SMAP and FINDSITE are stored in the sample-output directory; they should be used to verify the installation and working of these two programs.

Following, we provide detailed description on how to run Findsite, SMAP and ILbind. The **test.sh** script includes code to run the abovementioned programs. This script should be used to test the ILbind installation. In order to run the test, the first few lines in the test.sh script should be modified; the user should provide a path to binaries of SMAP and Findsite and the location of Findsite threading library located on the user's computer.

To verify the installation using our example, the outputs generated by test.sh that are located in the test-output folder should be compared against the contents of the sample-output folder.

Running FINDSITE

In the inverse binding prediction the Findsite's threading step is not required, as the templates with the ligand are already provided. FINDSITE should be executed as follows:

./findsite -s {PDB_ID}.pdb -m 0 -t templateFile -o {PDB_ID}

where

{PDB_ID}.pdb is(are) the target PDB structure(s) (Note that we recommend that all non-AA molecules are removed in these files); our example includes 1cqi_b, 1c30_c, 1ce8_c, 2c31_b, 2hmf_c, 1h2c_a, 1wpa_a, 1qdd_a, 1g3p_a, 2grr_b

templateFile is the file which list all considered templates; in our example the file is templates/ADP.tmpl

During the subsequent use, the whole PDB or its subset of a particular interest (e.g. human proteins) should be used.

Note that environmental variables in FINDSITE should be set up first; see FINDSITE's manual for more detailed instruction or investigate the test.sh script.

The templates may have more than one ligand. The user should note the name(s) of the ligand(s), which is/are stored in the Findsite database. In our example they are:

```
tmp1    2ZPAA01
tmp2    1GZFC01
tmp3    1CQIB01
```

A file with the template to ligand mapping must be created; an example file is templates/ADP.pairs

The file named {PDB_ID}.pockets.dat holds the information about the alignment length, which is generated by FINDSITE and used by ILbind as the input. ILbind prediction script will extract this information from the output files generated by FINDSITE; the script automatically selects the maximal alignment length across the templates. A more detailed description how to find this information (if user would like to manually check these values) follows:

Using our example, for 1c30_c.pockets.dat file (which is outputted by FINDSITE), the second line gives information about alignment to the 2zpa_a template:

```
TEMPLATE     1 2ZPAA   648   364    0.219    0.074    0.242    7.808
```

Alignment length is the 5th field (364). The results for all test proteins included in our example are as follows

	tmp1	tmp2	tmp3
1cqi_b	237	100	**385**
1c30_c	**364**	136	217
1ce8_c	**352**	133	234
2c31_b	**274**	126	185
2hmf_c	**257**	104	197
1h2c_a	**72**	67	**72**
1wpa_a	62	**72**	0
1qdd_a	0	**77**	0
1g3p_a	0	**208**	0
2grr_b	0	**89**	0

0 - means that protein was not aligned against a given template.

Bolded results (in a given row) represent the maximal alignment for a given target.

Running SMAP

SMAP should be run using the following command:

./smap_comp.sh {tmpl} {PDB_ID} {tmpl}_{PDB_ID}.smap

where

{tmpl} is the template used for the alignment; in our example it's one of the following: tmp1, tmp2, tmp3

{PDB_ID} is(are) the target PDB structure(s); this is similar as for FINDSITE, except the chain identifier is not included: 1cqi, 1c30, 1ce8, 2c31, 2hmf, 1h2c, 1wpa, 1qdd, 1g3p, 2grr

Note that both tempates and target protein(s) have to be gziped and copied to SMAP's PDB directory specified in SMAP's configuration file before running the program; both gziped input and template files, which are available in this distribution, must be copied to the PDB folder for test.sh to run correctly. A proper name of the SMAP file is {NAME}.pdb.gz where NAME is a 4 letter long pdb id or name of template.

User must prepare template files by removing all atoms that do not belong to the protein chain and the bound selected ligand. For the target protein(s), the user should use only one chain at the time and remove all atoms which are not part of the amino acid chain. The template and target protein(s) should be stored in files with names that do not contain the chain identifier. Moreover, the template file(s) should have a name that is different than the name of any of the files for the target proteins (e.g., tmp3 is template derived from 1cqi protein, and it should not be named 1cqi).

The output files *.smap include the raw score values, which are used as the input to ILbind. These values are extracted automatically by the ILBind script. Here, we provide information how to find these values manually:

Using our example, tmp1_1c30_c.smap file (which is outputted by SMAP) includes:

```
>Query Chain: 1C30_C    Query Ligand: ???       Template Ligand: ADP

P-Value = 2.929E-03      Raw Score = 52.02        Tanimoto Coeff = 0.11    Target Cover =
0.6      Query Cover = 0.14      RMSD = 2.4A
```

The raw score is 52.02. The results for all test proteins included in our example are as follows

	tmp1	tmp2	tmp3
1cqi_b	50.66	27.71	**269.94**
1c30_c	52.02	34.12	**70.4**
1ce8_c	52.62	34.43	**64.49**
2c31_b	40.39	28.88	**44.15**
2hmf_c	**62.91**	42.58	38.56
1h2c_a	23.9	17.42	**27.94**
1wpa_a	**25.7**	17.06	22.1
1qdd_a	26.04	21.24	**26.3**
1g3p_a	**37.59**	22.83	29.66
2grr_b	**32.12**	31.21	28.22

Running ILbind

The final prediction is performed using the ILbind.sh file. For each protein and template set, the user should run the following command:

./predauto.sh {PDB_ID} templates.lst resultsDir {tmpDir}

where

{PDB_ID} is(are) the target PDB structure(s); our example includes 1cqi_b, 1c30_c, 1ce8_c, 2c31_b, 2hmf_c, 1h2c_a, 1wpa_a, 1qdd_a, 1g3p_a, 2grr_b

template.pairs is a list of templates and corresponding ligand ids; see templates/ADP.pairs

resultsDir points to the locations of the SMAP (resultsDir/{template}_{PDB_ID}.smap) and Findsite (resultsDir/{PDB_ID}.pockets.dat) predictions.

{tmpDir} is used to modify the default temp directory (the directory where temporary files are stored). This directory will be deleted before program finishes, which is useful in case when ILbind is used concurrently.

The script outputs probabilities for each SVM model used by ILbind and, in the last line, the overall predicted probability of binding (calculated as an average probability from the fifteen SVM models). Higher probability indicates higher chances that the considered ligand will bind to a given protein.

Using our example, for the 1c30_c target:

./ILbind.sh 1c30_c templates/ADP.pairs sample-output/1c30_c

where

templates/ADP.pairs has following content:

```
tmp1  2ZPAA01
tmp2  1GZFC01
tmp3  1CQIB01
```

sample-output/1c30_c folder contains following files:

- 1c30_c.pockets.dat
- tmp1_1c30_c.smap
- tmp2_1c30_c.smap
- tmp3_1c30_c.smap

These results should be used to validate whether the installation was successful, i.e., whether the user generated outputs and the outputs provided with this distribution match.