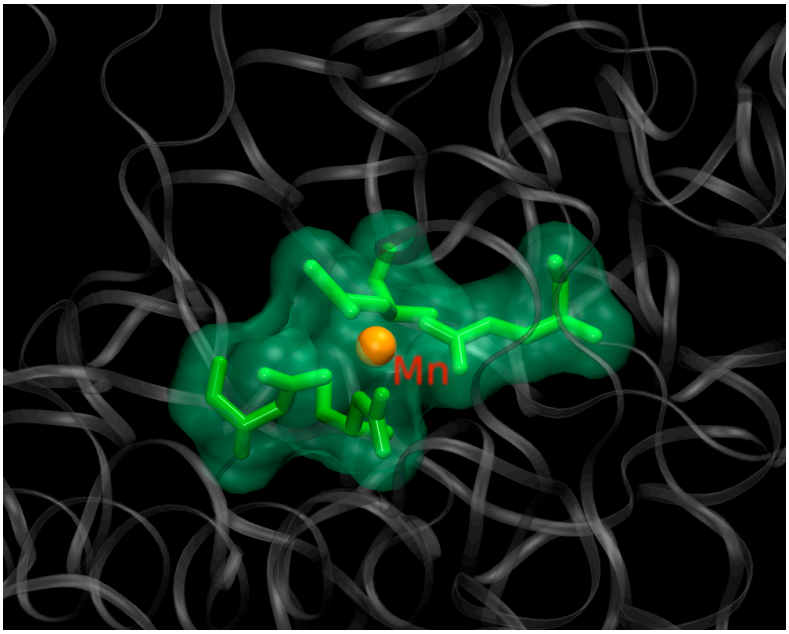


Version 1.0 | 2010

FINDSITE-metal manual

Michal Brylinski

Jeffrey Skolnick

Please report bugs and other issues to michal@gatech.edu

Table of Contents

1. Installation and requirements

1.1. Required libraries

Before you start compiling FINDSITE-metal, make sure that the following libraries are available on your system:

zlib	http://www.zlib.net/
gzstream	http://www.cs.unc.edu/Research/compgeom/gzstream/
f2c	http://www.netlib.org/f2c/
libsvm	http://www.csie.ntu.edu.tw/~cjlin/libsvm/

If you have them installed in locations not recognized by your C++ compiler, depending on your configuration, you may want to copy following files to findsite-1.0/lib/:

zlib	zlib.h, zconf.h, libz.a
gzstream	gzstream.h, libgzstream.a
f2c	f2c.h, libf2c.a
libsvm	svm.h, libsvm.a

To generate libsvm.a, you can either use a patch that is in findsitemetal-1.0/patch/:

```
[local]$ tar -xzf libsvm-2.9.tar.gz
[local]$ cd libsvm-2.9/
[libsvm-2.9]$ patch < ../findsitemetal-1.0/patch/libsvm-
2.9.patch
[libsvm-2.9]$ make
[libsvm-2.9]$ cp svm.h libsvm.a ../findsitemetal-1.0/lib/
```

or simply use ar:

```
[local]$ tar -xzf libsvm-2.9.tar.gz
[local]$ cd libsvm-2.9/
[libsvm-2.9]$ make
[libsvm-2.9]$ ar cr libsvm.a svm.o
[libsvm-2.9]$ cp svm.h libsvm.a ../findsitemetal-1.0/lib/
```

Now go to findsite-1.0/src/ and run:

```
[src]$ make
```

Note that this step requires some of the Perl modules described below, so get them installed before you run make. If there are no problems, the following files will show up in findsitemetal-1.0/bin/:

```
[src]$ ls ../bin/
findsite_function
findsite_library
findsitemetal
findsitemetal_conf
generate_conf_files
```

1.2. Perl modules

You will need following Perl modules that are available from CPAN:

Algorithm::Numerical::Shuffle	http://search.cpan.org/~abigail/Algorithm-Numerical-Shuffle-2009110301/
AI::Calibrate	http://search.cpan.org/~tomfa/AI-Calibrate-1.1/

AI::NaiveBayes1	http://search.cpan.org/~vlado/AI-NaiveBayes1-1.8/
File::Slurp	http://search.cpan.org/~drolsky/File-Slurp-9999.13/
File::Temp	http://search.cpan.org/~tjenness/File-Temp-0.22/
GO::TermFinder	http://search.cpan.org/~sherlock/GO-TermFinder-0.86/
Math::Counting	http://search.cpan.org/~gene/Math-Counting-0.0801/
Uniq	http://search.cpan.org/~syamal/Uniq-0.01/
YAML	http://search.cpan.org/~adamk/YAML-0.71/

1.3. NCBI BLAST

Finally, you should have NCBI BLAST installed, configured and available from the default search path. To verify this, please run:

```
[src]$ which formatdb blastall
/opt/local/bin/formatdb
/opt/local/bin/blastall
```

Of course, /opt/local/bin/ may be different on your system. NCBI BLAST can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/.

2. Environmental variables

Before you run FINDSITE-metal please set the following variables:

```
[src]$ export FINDSITELIB=/some_path/pdb_library
[src]$ export FINDSITEMAP=/some_path/my_map_file
[src]$ export FINDSITEDAT=/some_path/findsitemetal-1.0/dat
```

```
[src]$ export GOOBO=/some_path/gene_ontology.1_2.obo
[src]$ export GOPDB=/some_path/gene_association.goa_pdb
[src]$ export GOFRQ=/some_path/UniProt_090109.frq
```

Replace “some_path” with the correct path, see example Section 5.3.

See Section 4 for library files required to set FINDSITELIB and FINDSITEMAP.

FINDSITEDAT should point at the findsitemetal-1.0/dat directory that contains following files (YAML and calibration files are created by make):

- findsitemetalNB1.cal
- findsitemetalNB1.yaml
- findsitemetalNB2.cal
- findsitemetalNB2.yaml
- findsitemetalNB3.cal
- findsitemetalNB3.yaml
- findsitemetalSVM.model
- findsitemetalSVM.scale

You can find UniProt_090109.frq in findsitemetal-1.0/dat/. The files gene_ontology.1_2.obo and gene_association.goa_pdb can be downloaded from http://www.geneontology.org/GO.downloads.shtml and ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/PDB/, respectively.

3. Programs

The FINDSITE-metal software distribution includes the following programs (to get the list of available options for each program, execute it with no arguments):

Name	**findsite_library**
Description	prepares a library map file for use by FINDSITE/FINDSITE-metal
ENV	
Requirements	formatdb, blastall, File::Slurp, File::Temp
Input files	• template library in FASTA format • PDB library (see Section 4)
Output files	map file for FINDSITE/FINDSITE-metal
Mandatory arguments	-t template library in FASTA format -p PDB library -o output filename
Optional arguments	

Name	**findsitemetal**	
Description	main FINDSITE-metal program that predicts metal-binding residues and metal preferences	
ENV	FINDSITELIB, FINDSITEMAP, FINDSITEDAT	
Requirements	zlib, gzstream, f2c, libsvm	
Input files	• target protein structure in PDB format, only one chain is allowed • list of templates, e.g. identified by threading	
Output files	*.sites.dat	detailed information on the detected metal-binding sites
	*.sites.pdb	geometric centers of the detected sites, PDB format
	*.alignments.dat	structure alignments from fr-TM-align
	*.metals.pdb	metals extracted from the templates, PDB format
	*.templates.pdb	metal-bound templates aligned onto the target structure using fr-TM-align
Mandatory	-s target structure in PDB format	

arguments	-t	list of templates
	-o	output filename
Optional arguments	-b	max sequence identity of templates to the target, default 1.0
	-m	TMscore threshold, default 0.4
	-r	SVM probability threshold for binding residues, default 0.15
	-n	min number of binding residues, default 1
	-c	estimated TM-score of the target structure to native, default 1.0 (see Appendix A)
	-l	TM-score normalization, default T • T – by the target length • S – by the length of a smaller protein

Name	**findsite_function**
Description	assigns Gene Ontology terms to the predicted metal-binding sites (also works for ligand-binding pockets predicted by FINDSITE)
ENV	GOOBO, GOPDB, GOFRQ
Requirements	Algorithm::Numerical::Shuffle, File::Slurp, GO::TermFinder, Math::Counting, Uniq
Input files	• metal-binding sites identified by FINDSITE-metal
Output files	list of GO terms assigned to the metal-binding sites
Mandatory arguments	-p FINDSITE-metal sites -o output filename
Optional arguments	-t probability threshold for function transfer, default 0.5

Name	**findsitemetal_conf**
Description	estimates the prediction confidence for FINDSITE-metal
ENV	FINDSITEDAT
Requirements	AI::NaiveBayes1, File::Slurp, YAML

Input files	• metal-binding sites predicted by FINDSITE-metal
Output files	confidence estimates
Mandatory arguments	-s FINDSITE-metal sites
Optional arguments	-c estimated TM-score of the target structure to native, default 1.0 (see Appendix A)

Name	**generate_conf_files**
Description	generates files required by the prediction confidence, this script is executed by make
ENV	
Requirements	AI::Calibrate, AI::NaiveBayes1, File::Slurp, YAML
Input files	
Output files	findsitemetalNB1.cal, findsitemetalNB1.yaml, findsitemetalNB2.cal, findsitemetalNB2.yaml, findsitemetalNB3.cal, findsitemetalNB3.yaml
Mandatory arguments	
Optional arguments	

4. PDB library

In order to run FINDSITE/FINDSITE-metal, you will need a PDB library that can be downloaded from http://cssb.biology.gatech.edu/findsite/ and a mapping file, which maps the template library to the PDB (see ref 5). It can be created by **findsite_library** (see example, Section 5.1.). For a given template library (e.g. used in threading), you need to generate the mapping file only once. As long as you don't change your template library, you don't have to create new mapping files. If you have the mapping file already generated for FINDSITE, you can use it for FINDSITE-metal as well.

5. Example

Go to findsitemetal-1.0/example/, where you will find following files:

- **1a5vA.pdb** – crystal structure of our target, HIV-1 integrase (PDB-ID: 1a5v, chain A)
- **1a5vA.templates.lst** – a list of templates identified for the target sequence (13gsA) by threading
- **my_template_library.fasta** – a complete template library used in threading (FASTA format)
- **sample_pdb_library** – a sample PDB library. **Do not use it outside this example.** Instead, download the latest version from http://cssb.biology.gatech.edu/findsite/

5.1. Mapping file for FINDSITE/FINDSITE-metal

To prepare the mapping file for FINDSITE/FINDSITE-metal run:

```
[example]$ perl ../bin/findsite_library \
-t my_template_library.fasta \
-p sample_pdb_library/PROTEINS.FAS \
-o my_mapping.cls
```

After some time, you should get the my_mapping.cls file that contains PDB entries mapped to your templates. You can use this file for both FINDSITE and FINDSITE-metal.

5.2. Template structure identification

Run threading (see refs 7-10) or a similar procedure to identify template structures for your target protein. In principle, any sequence profile-based approach (e.g. HHpred, ref 11) should work (see ref 4). FINDSITE-metal requires two input files: target protein structure in PDB format (crystal structures as well as protein models can be used) and the list of identified template structures. In our example, the input files are: 1a5vA.pdb

and 1a5vA.templates.lst. The latter was obtained from PROSPECTOR_3/SP3/SPARKS2 (see refs 8-10).

5.3. Environmental variables

Set several environmental variables pointing at the locations of the PDB library (FINDSITELIB), the mapping file (FINDSITEMAP), FINDSITE-metal data directory (FINDSITEDAT) and the Gene Ontology files (GOOBO, GOPDB and GOFRQ). In this example make exports as follows:

```
[example]$ export FINDSITELIB=sample_pdb_library
[example]$ export FINDSITEMAP=my_mapping.cls
[example]$ export FINDSITEDAT=../dat
[example]$ export GOOBO=../dat/gene_ontology.1_2.obo
[example]$ export GOPDB=../dat/gene_association.goa_pdb
[example]$ export GOFRQ=../dat/UniProt_090109.frq
```

5.4. FINDSITE-metal

Run FINDSITE-metal with no arguments to get a list of available options:

```
[example]$ ../bin/findsitemetal
```

Now run FINDSITE-metal for the target, 1a5vA, with default parameters:

```
[example]$ ../bin/findsitemetal \
-s 1a5vA.pdb \
-t 1a5vA.templates.lst \
-o 1a5vA.findsitemetal
```

Since the target crystal structure is used, we could use the default value of the TM-score (1.0). As the result you should get following output files:

- 1a5vA.findsitemetal.alignments.dat
- 1a5vA.findsitemetal.metals.pdb
- 1a5vA.findsitemetal.sites.dat
- 1a5vA.findsitemetal.sites.pdb
- 1a5vA.findsitemetal.templates.pdb

See Section 6 for output files. See Appendix A for info on how to estimate the TM-score if a protein model is used as the target structure.

5.5. Molecular function prediction

To predict the molecular function according to the Gene Ontology classification (see ref 12), use the following command:

```
[example]$ perl ../bin/findsite_function \
-p 1a5vA.findsitemetal.sites.dat \
-o 1a5vA.findsitemetal.function.out
```

Find the GO terms predicted for each pocket in 1a5vA.findsite.function.out.

See Section 6 for output files.

5.6. Prediction confidence

You can estimate how confident the prediction is using **findsitemetal_conf**:

```
[example]$ perl ../bin/findsitemetal_conf \
-s 1a5vA.findsitemetal.sites.dat
```

See Section 6 for the output format.

6. Output files

Take a look at the output files generated by FINDSITE-metal for the example above, 1a5vA.

1a5vA.findsitemetal.sites.dat. This file contains detailed information on the detected metal-binding sites. Sites are separated by the TER keyword. The content is divided into following sections:

SITE. This section contains some general information.

```
SITE      1   11  0.6471  1.4810  0.6243  0.1114    2
```

- site number (1)
- number of template-bound metals used to identify this site (11)
- fraction of templates that share this site (0.6471)
- sequence profile score (1.4810).
- average TM-score of the templates to the target structure (0.6243)
- standard deviation for the average TM-score (0.1114)
- number of predicted metal-binding residues (2)

TEMPLATE. List of template structures and their similarities to the target.

```
TEMPLATE   1 1BIUA  147  124   0.325  0.250  0.762  1.876
```

- template number (1)
- template PDB-ID (1BIUA)
- number of residues (147)

- alignment length – number of residues aligned to the target by fr-TM-align (124)
- global sequence identity to the target (0.325)
- sequence identity calculated over the residues aligned by fr-TM-align (0.250)
- TM-score to the target structure (0.762)
- Cα-RMSD [Å] of the aligned region (1.876)

METAL. Calculated preferences for binding metals.

```
METAL CA 0.000000 0.059177
```

- metal ID (CA)
- binding probability (0.000000)
- binding probability that accounts for the background binding frequencies of metal ions in the PDB (0.059177)

MENTROPY. Shannon's entropy for metal-binding probabilities.

```
MENTROPY    1.241    1.913    2.583
```

- entropy (1.241)
- entropy that accounts for the background binding probabilities (1.913)
- maximum entropy (2.583)

CENTER. Coordinates of the predicted site center (x, y, z).

```
CENTER   52.780  37.713  58.366    1.450   1.935   0.848
```

- site center *x*, *y*, *z* (52.780, 37.713, 58.366)
- average *x*, *y*, *z* displacement (1.450, 1.935, 0.848)

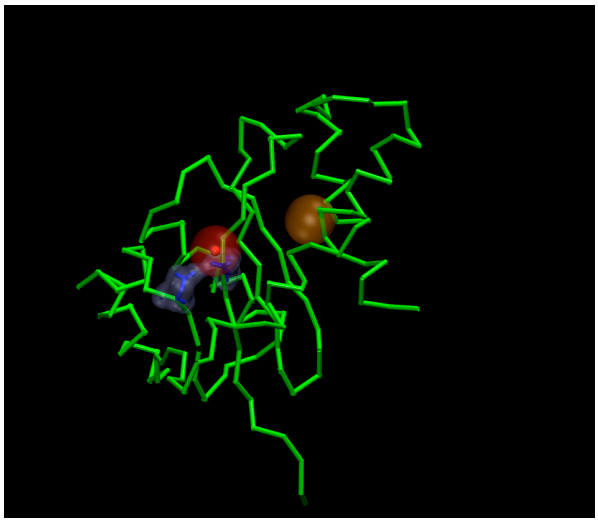
RESIDUE. List of predicted metal-binding residues.

```
RESIDUE 11 D * 7  4.577 0.8198 0.5014 1.0000 0.7405 0.3755
```

- residue number (11)
- residue 1-letter code (D)
- asterisks mark residues likely to bind metals (*)
- number of templates that have a residue in equivalent position in contact with a metal (7)
- distance between residue Cα atom and the predicted site center (4.577)
- metal-binding probability estimated by SVM (0.8198)
- metal-binding probability (fraction of templates) that accounts for the background binding frequencies of metals in the PDB (0.5014)
- sequence profile score calculated from the set of metal-bound templates (1.0000)
- sequence profile score that accounts for the background metal-binding preferences of amino acids (0.7405)
- generic preference to bind the top-ranked predicted metal (0.3755). In this case, the top-ranked metal is MG with the probability of 0.521926

1a5vA.findsitemetal.sites.pdb. This file contains the geometric centers of the predicted and ranked metal-binding sites in PDB format.

```
HETATM    1  FS  PKT M  11      52.780  37.713  58.366
HETATM    2  FS  PKT M   3      54.418  34.614  46.314
```



The atom number and the residue number are replaced with the site rank and the number of metal ions used to identify this site, respectively. You can display this file along with your target structure to see the locations of the predicted metal-binding sites. In the figure on the left, the top-ranked binding site (binding residues) is shown in red (blue), whereas a site on the lower rank is colored in brown. The target structure is represented as the green Cα trace.

1av5.findsitemetal.alignments.dat. This file contains structure alignments generated by fr-TM-align (see ref 13).

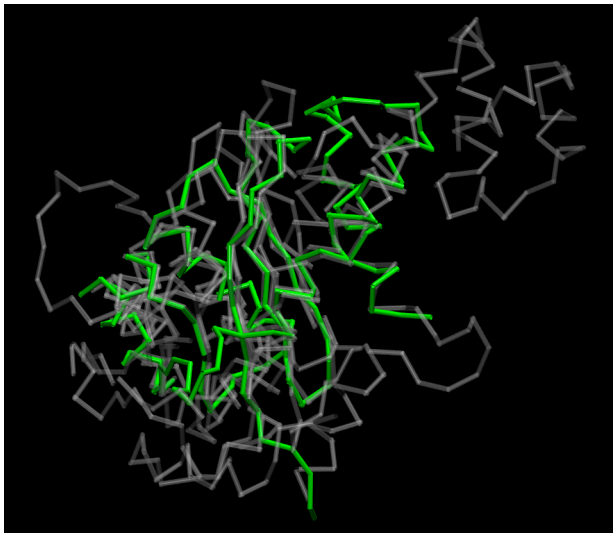
```
>1BIUA 147 124 0.762 1.876 0.250
GLGPLQIWQTDFTLEPRMAPRSWLAVTVDTASSAIVVTQHGRVTSVAAQHHWATAIAV
  .:::::::::::::  .  ::::::::::::::::::::::::::::::::::::
--CSPGIWQLDCTHLE--G--KVILVAVHVASGYIEAEVIPAETGQETAYFLLKLAGR
*
```

- the first line reports the PDB-ID of a template (1BIUA), template length (147), alignment length (124), TM-score to the target structure (0.762), Cα-RMSD [Å] calculated over the aligned residues (1.876) and the sequence identity of the aligned residues (0.250)
- the second and fourth lines show the aligned sequences of the target and the template, respectively
- the third line highlights the aligned residue positions (.) and the residue pairs, whose Cα atoms are aligned within a distance of 5 Å (:)

- * separates alignments for different templates

1a5vA.findsitemetal.templates.pdb. In this file, you will find template structures in PDB format aligned onto the target structure by fr-TM-align.

```
REMARK    PDB-ID:      1BIUA
REMARK    SEQID1:      0.325
REMARK    SEQID2:      0.250
REMARK    TM-SCORE:    0.762
REMARK    RMSD:        1.876
```

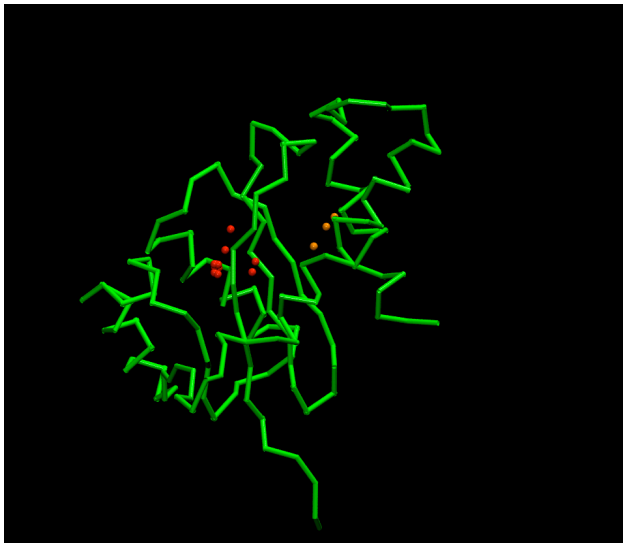


The *REMARK* section reports the PDB-ID of the template (1BIUA), global sequence identity to the target (0.325), sequence identity calculated over the residues aligned by fr-TM-align (0.250), TM-score to the target structure (0.762) and Cα-RMSD [Å] of the aligned region (1.876). Please note that the templates are aligned onto the target structure; this is shown in the figure above (green – target, gray – selected templates).

1a5vA.findsitemetal.metals.pdb. This PDB file contains the metal ions extracted from the templates upon superposition onto the target structure by fr-TM-align (see ref 13).

```
HETATM    1 MG    MG      1       54.328  37.819  58.428
HETATM    2 MG    MG      1       51.231  38.712  58.930
HETATM    3 MN    MN      1       51.193  38.273  58.833
```

```
HETATM    4 MG    MG      1       51.772  38.965  58.679
HETATM    5 MG    MG      1       51.536  38.826  58.221
HETATM    6 MN    MN      1       53.392  33.550  58.209
HETATM    7 MN    MN      1       55.729  39.163  56.494
HETATM    8 MG    MG      1       52.533  38.154  59.013
HETATM    9 MG    MG      1       52.208  38.738  58.734
HETATM   10 MG    MG      1       54.069  34.229  57.142
HETATM   11 ZN    ZN      1       52.585  38.412  59.341
TER
HETATM   12 MN    MN      2       55.305  35.048  44.541
HETATM   13 ZN    ZN      2       53.391  33.897  48.552
HETATM   14 MG    MG      2       54.560  34.896  45.847
TER
```



The residue number is replaced with the site rank. Sites are separated by the TER keyword. You can display this file along with your target structure to see the locations of the template-bound metal ions. In the figure on the left, template-bound metals that form the top-ranked binding site are shown in red, whereas these that form a site at the lower rank are colored in orange. The target structure is presented as a green Cα trace.

1a5vA.findsitemetal.function.out. List of Gene Ontology terms assigned to the metal-binding sites predicted by FINDSITE-metal.

```
SITE    1 GO:0046914  1.000000  0.501172   1.628e-01
```

- site number (1)
- GO molecular function term (GO:0046914)
- function transfer probability (1.000000)
- function transfer probability that accounts for the background frequencies of GO terms in UniProt (0.501172)
- *p*-value calculated using Fisher's exact test (1.628e-01)

Confidence estimates. Three confidence estimates are calculated:

```
Prediction confidence for the estimated TMscore of 1.000:

Distance <= 4 A     0.774    63.8%
Residues >= 0.50    0.726    66.7%
Metal               0.425    50.0%
```

- chances that the top-ranked site is predicted within a distance of 4Å (63.8%)
- chances that the Matthew's correlation coefficient of the predicted binding residues is ≥0.50 (66.7%)
- chances that the top-ranked metal is correctly predicted (50.0%)

7. References

Primary citation for FINDSITE-metal

[1] Brylinski M, Skolnick J (2010) FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal binding site prediction at the proteome level. *Submitted*.

FINDSITE references

[2] Brylinski M, Skolnick J (2010) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci USA* **105**: 129-134.

[3] Skolnick J, Brylinski M (2009) FINDSITE: a combined evolution/structure-based approach to protein function prediction. *Brief Bioinform* **10**: 378-391.

[4] Brylinski M, Skolnick J (2009) Comparison of structure-based and threading-based approaches to protein functional annotation. *Proteins* **78**: 18-134.

Other references

[5] Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C (2002) The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **58**: 899-907.

[6] Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**: 27-30.

[7] Jones DT, Hadley C (2000) Threading methods for protein structure prediction. In: Higgins D, Taylor WR, editors. Bioinformatics: sequence, structure and databanks. *Heidelberg: Springer-Verlag*; pp 1-13.

[8] Skolnick J, Kihara D, Zhang Y (2004) Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins* **56**: 502-518.

[9] Zhou H, Zhou Y (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* **55**: 1005-1013.

[10] Zhou H, Zhou Y (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* **58**: 321-328.

[11] Söding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**: 951-960.

[12] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet* **25**: 25-29.

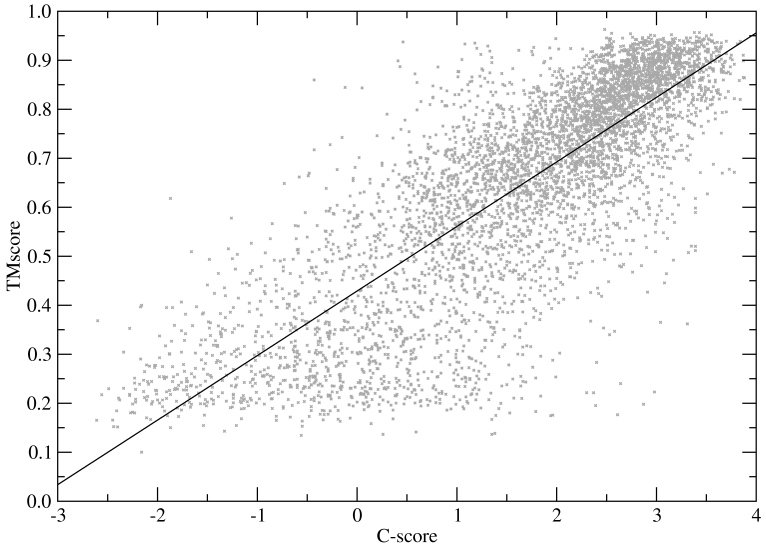
[13] Pandit SB, Skolnick J (2008) Fr-TM-align: A new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics* **9**: 531.

[14] Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. *Proteins* **57**: 702-710.

[15] Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci U S A* **101**: 7594-7599.

Appendix A – Estimation of the TM-score for protein models

TM-score provides a length-independent measure of structural similarity between two proteins (ref. 14). A significant similarity is indicated by a TM-score of >0.4. FINDSITE-metal uses the estimated TM-score to native as one of the SVM features to accurately predict metal-binding residues. For a protein model, the TM-score can be directly calculated against the native crystal structure, if known. However, in a real scenario, when the experimental structure of a target is not available, the TM-score needs to be estimated. Most of the contemporary structure prediction algorithms estimate the reliability of structure modeling using some score. TASSER, a template-based structure assembly/refinement approach (ref. 15), calculates a confidence score, called C-score.



To estimate the TM-score for target structures modeled by TASSER, we carried out large-scale benchmarks on a non-redundant and representative dataset of protein targets (see ref. 1). For each modeled target structure, we calculate the TM-score vs. its crystal structure and plot it against a C-score obtained from TASSER simulations. This is shown in a figure above. Next, we calculate the regression line, which in this case is:

$$TM - score = 0.13173 \times C - score + 0.42895$$

We use this equation to estimate the TM-score for target structures modeled by TASSER. FINDSITE-metal does not require the exact TM-score; an estimate works fine.