9-29-2009

# EASYMIFS and SITEHOUND: a toolkit for the identification of ligand-binding sites in protein structures

Dario Ghersi
*University of Nebraska at Omaha*, dghersi@unomaha.edu

Roberto Sanchez
*Mount Sinai School of Medicine*

# EASYMIFS and SITEHOUND: a toolkit for the identification of ligand-binding sites in protein structures

Dario Ghersi and Roberto Sanchez

**Abstract**

**Summary:** SITEHOUND uses Molecular Interaction Fields (MIFs) produced by EASYMIFS to identify protein structure regions that show a high propensity for interaction with ligands. The type of binding site identified depends on the probe atom used in the MIF calculation. The input to EASYMIFS is a PDB file of a protein structure; the output MIF serves as input to SITEHOUND, which in turn produces a list of putative binding sites. Extensive testing of SITEHOUND for the detection of binding sites for drug-like molecules and phosphorylated ligands has been carried out.

**Availability:** EASYMIFS and SITEHOUND executables for Linux, Mac OS X, and MS Windows operating systems are freely available for download from http://sitehound.sanchezlab.org/download.html.

**Contact:** roberto@sanchezlab.org or roberto.sanchez@mssm.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1. Introduction

The molecular function of a protein is largely determined by interactions with other molecules at binding sites on its surface. Hence, identification of the location and characteristics of ligand-binding sites can contribute to functional annotation of a protein; it can guide experiments, and be useful in predicting or verifying interactions. The identification of ligand-binding sites can also be an important part of the drug discovery process.

Several methods have been developed for the identification of binding sites from protein structures and sequences. Sequence-based methods (Berezin *et al.*, 2004; Capra and Singh, 2007) have the advantage of being applicable to proteins of unknown structure, by relying on the evolutionary conservation of residues. However, they are also limited by the fact that not all binding sites are conserved, and not all conserved residues correspond to binding sites. Structure-based approaches can overcome these limitations and complement sequence-based methods (Wang *et al.*, 2008). The simplest example of this complementarity is the use of a protein structure as a filter for sequence-based predictions (Lichtarge and Sowa, 2002). Some structure-based methods rely on geometrical features (Dundas *et al.*, 2006), while others rely on energetic calculations. Energy-based methods can identify conserved residue side-chains in high-energy conformations, which tend to correspond to functionally relevant residues (Cheng *et al.*, 2005; Liang *et al.*, 2006) or directly describe the molecular interaction properties of the protein surface (Laurie and Jackson, 2005). Hence, these methods can identify non-conserved binding sites, and identify residues that are conserved solely due to functional constraints. The methods that rely on molecular interaction properties can in principle also distinguish binding sites with distinct ligand preferences (e.g. hydrophobic versus polar) if different chemical probes are used for the molecular interaction calculation.

Here we describe software tools that enable the identification of binding sites via calculation of molecular interaction properties. EASYMIFS, a simple Molecular Interaction Field (MIF) calculator; and SITEHOUND, a post processing tool for MIFs that identifies interaction energy clusters corresponding to putative binding sites. The motivation for the development of these tools is to provide easy to use and freely available software for the analysis of protein structures on a wide range of computer systems. While a web interface to SITEHOUND is available (Hernandez *et al.*, 2009), it is limited to smaller systems and only two types of binding sites. The EASYMIFS and SITEHOUND tools provide a large set of chemical probes for MIF calculation and are designed to facilitate large scale and automated analysis, for which web interfaces are not practical.

## 2.0 Methods

### 2.1 MIF calculation

MIFs describe the spatial variation of the interaction energy between a target molecule (e.g. a protein) and a probe, which represents a specific chemical group or atom. The MIF is usually discretized on a 3D orthogonal grid that surrounds the target molecule, thus representing a map of the potential energy between the probe and the target molecule. Some of the applications of MIFs include quantitative structure–activity relationship (QSAR), ligand selectivity analysis, pharmacophoric search and binding site detection (Cruciani, 2006). EASYMIFS provides a simple and rapid way of characterizing a protein structure from a chemical standpoint, returning maps that can be displayed in molecular graphics software such as PyMOL (DeLano) and others. The MIF calculations are carried out *in vacuo* using the GROMOS force field (Van Der Spoel *et al.*, 2005) and a distance dependent dielectric. For each grid point in the discretized orthogonal grid, the program computes the potential energy between the probe (a particular atom type) and all atoms in the protein, as a sum of van der Waals and electrostatic interactions. The parameters for the Lennard-Jones term, representing the van der Waals interactions, depend on the chosen probe and the particular protein atom type, and are taken from a matrix of parameters distributed with the GROMACS package (Van Der Spoel *et al.*, 2005). The distance-dependent dielectric sigmoidal function of Solmajer and Mehler (1991) as described in Cui *et al.* (2008) is used. The default output is a 1 Å resolution grid with dimensions large enough to accommodate the entire protein.

### 2.2 Binding site identification

The SITEHOUND algorithm identifies the location of potential binding sites by filtering and clustering the MIFs produced by EASYMIFS or other programs in a way similar to Q-SiteFinder (Laurie and Jackson, 2005). The main differences lie in the use of multiple probes for the detection of different types of binding sites; alternative clustering algorithms, which improve results for ligands of different shapes (see Supplementary Materials); and the fact that SITEHOUND can be run independently of a web interface. In the first step, a probe-dependent energy cutoff is applied to filter out all the MIF points corresponding to unfavorable interaction energies. The remaining points are clustered according to spatial proximity with an agglomerative hierarchical clustering algorithm using average or single linkage. This step yields a hierarchical dendrogram, which is cut into non-overlapping clusters by applying a distance cut-off. Total interaction energy (TIE; the sum of the energy of all the points in the cluster) is used to rank the resulting clusters. A test on 77 protein–ligand complexes containing drug-like molecule

showed that the correct site is identified among the top three SITEHOUND clusters in 95% of the cases (79% for unbound proteins) when using the 'methyl' probe (Ghersi and Sanchez, 2009). Similar accuracy was observed in a set of more than 120 proteins that bind to phosphorylated ligands when using the 'phosphate oxygen' probe for binding site detection (Ghersi and Sanchez, manuscript in preparation).

## 2.3 Usage and output

While EASYMIFS and SITEHOUND can be used separately (see Supplementary Materials) they are usually used in combination with each other. A script called auto.py allows for fully automated binding site identification using a single command: '**auto.py –i 1kna.pdb –p PROBE**'. Where the **–i**option specifies the input PDB file (e.g. **1kna.pdb**) and the **–p** option specifies the probe used for binding site detection. More than 30 atom types from the GROMOS forcefield are available to use as probes. As mentioned above, the methyl (CMET) and phosphate oxygen (OP) probes have been tested extensively. Additionally, hydroxyl oxygen (OA) has been used to characterize sugar-binding sites; and peptide nitrogen (N), water (OW) and bare carbon (C) have been used to characterize various other ligand binding sites. The dimensions of a box large enough to enclose the whole protein, with a clearance of 5 Å in each direction and a resolution of 1 Å are determined automatically. Alternatively, command line options can be used to specify the center, dimensions and resolution of the grid. Several output files are produced, the most important ones being:_summary.dat that contains a summary of all the clusters ranked by TIE;_predicted.dat that lists the protein residues in contact with each of the clusters; and _clusters.pdb, which can be used to display the clusters in molecular graphics programs (Fig. 1). Calculations typically take a few minutes for one protein depending on the computer system and the size of the protein (see Supplementary Materials). Most of the computing time is taken up by the MIF calculation in EASYMIFS. The ability to run SITEHOUNDseparately from EASYMIFS allows the user to explore different clustering options (e.g. average or single linkage) and parameters (energy and distance cutoff) without the need to recalculate the MIF.
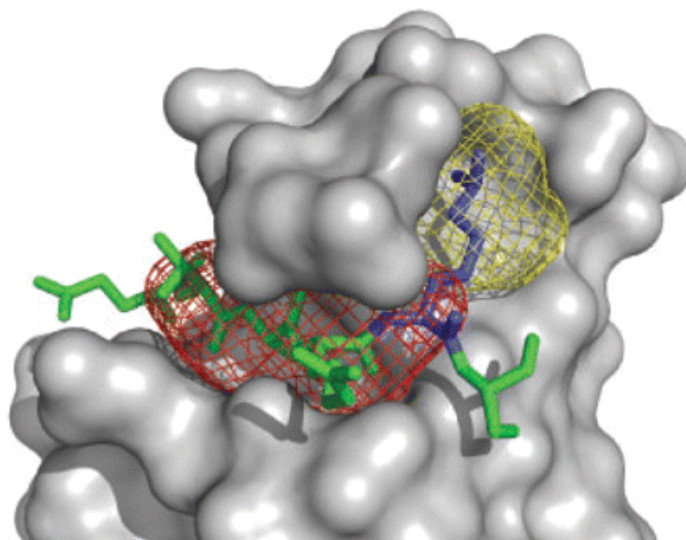


**Fig. 1.**

Sample SITEHOUND output. Chromo domain of HP1 complexed with histone H3 tail (green) containing dimethyl-lysine 9 (blue) (PDB 1kna). SITEHOUND methyl probe clusters are shown as mesh surfaces. The

red cluster identifies the peptide-binding region; the yellow cluster identifies the dimethyl-lysine-binding environment. The figure was rendered with PyMOL (DeLano) using the SITEHOUND '_clusters.pdb' output file.

# References

Berezin,C. et al. (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics, 20, 1322–1324.

Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. Bioinformatics, 23, 1875–1882.

Cheng,G. et al. (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res., 33, 5861–5867.

Cruciani,G. (2006) Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction. Wiley-VCH, Chichester.

Cui,M. et al. (2008) Prediction of protein loop structures using a local move Monte Carlo approach and a grid-based force field. Protein Eng. Des. Sel., 21, 729–735.

DeLano,W.L. The PyMOL Molecular Graphics System. Available at http://www.pymol.org.

Dundas,J. et al. (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res., 34, W116–W118.

Ghersi,D. and Sanchez,R. (2009) Improving accuracy and efficiency of blind proteinligand docking by focusing on predicted binding sites. Proteins, 74, 417–424.

Hernandez,M. et al. (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. Nucleic Acids Res., 37, W413–W416.

Laurie,A.T. and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics, 21, 1908–1916.

Liang,S. et al. (2006) Protein binding site prediction using an empirical scoring function. Nucleic Acids Res., 34, 3698–3707.

Lichtarge,O. and Sowa,M.E. (2002) Evolutionary predictions of binding surfaces and interactions. Curr. Opin. Struct. Biol., 12, 21–27.

Solmajer,T. and Mehler,E.L. (1991) Electrostatic screening in molecular dynamics simulations. Protein Eng., 4, 911–917.

Van Der Spoel,D. et al. (2005) GROMACS: fast, flexible, and free. J. Comput. Chem., 26, 1701–1718.

Wang,K. et al. (2008) Protein meta-functional signatures from combining sequence, structure, evolution, and amino acid property information. PLoS Comput. Biol., 4, e1000181.