SiteSeeker: an Algorithm to Identify Ligand Binding Sites in Proteins on a Genome Scale

1. Introduction

In the post-genome area, identification and characterization of ligand binding sites of proteins play increasing roles for drug discovery. Ligand binding site annotation may be used to identify and validate drug targets, to prioritize and optimize drug lead, to rationalize small molecule screening and docking, to guide medical chemistry effort to design druggable molecules, and to evaluate ADMET properties of preclinical drugs computationally. Therefore, a sensitive and robust algorithm that can identify and characterize the ligand binding site in the protein on a genome scale will be extremely valuable in chemical genomics studies and thus provides drug discovery processes with vital information.

Current methods for determining ligand binding sites, or equivalently, amino acid residues or clusters of residues include primary sequence comparison methods [], geometry methods based on the analysis of 3D structures [], and physical based methods that use physical or statistical energy functions []. All of these methods have their advantages and limitations. The sequence-based method requires a collection of diverse sequences. However, for a large amount of sequences, their homologous are limited. Even there are enough homologous sequences available, evolutionary signals may be too weak to be detected for a large amount of binding sites. The geometry-based methods are useful to detect deep pockets and cavities. However, it may miss shallow pockets such as protein-protein interaction sites. In addition, it is difficult to determine the boundary of predicted binding sites with the geometry-based methods. The physical properties of binding sites are critical to the ligand binding. By using molecular mechanics and simulation, physiochemical properties of the protein surface can be characterized, and applied to the prediction of the ligand binding sites. Unfortunately, the physical-based methods are usually computational-intensive, and difficult to be applied in a large scale. Therefore, to apply successfully the ligand binding site prediction to the chemical genomics project, the prediction algorithm has to satisfy the following requirements: First, it should identify both of the location and the boundary of binding site accurately without human intervention; Secondly, it should assess on the confidence of the predication more accurately. Third, it should be both sensitive and robust enough independent on protein families, and types and sizes of binding sites. Finally, it can be applied not only to experimental structures from X-ray or NMR, but also to theoretical models. Unfortunately, no current methods can satisfy all of above requirements.

In this paper, we present a new method to predict ligand binding sites of proteins. It is fast, sensitive and robust, thus can be applied to the large scale chemical genomics projects.

Methods

Overview of algorithms

There are several unique features of the SiteSeeker algorithm. (1) The SiteSeeker integrates evolutionary and geometry information of protein sequence and structure into a unified computational framework. (2) The SiteSeeker treats the protein in a case-by-case basis. For an interested protein, it attempts to find the most distinguished characteristics on its structure and sequence besides general considerations for the binding site. (3) The SiteSeeker dynamically adjusts parameters during the prediction and determines the boundary of predicted site accurately. (4) Finally, SiteSeeker provides an estimation of statistics significance on the prediction.

The SiteSeeker algorithm consists of three basic components. The first component determines which residues are the most unlikely to play roles in the functional sites based on evolution and geometry criteria. This kind of residues is termed as the rim residue. By combining information of the rim residue and the surface mesh triangle, the second component computes disjointed 2D surface patches that are candidates of binding sites. Finally, the third component constructs 3D pockets with the surface patches from the second component and the residues from the first component, and determines the boundary of the pocket automatically. The pocket's geometry properties such as volume, surface area, depth, and mouth area are calculated after the pocket is defined. The final predictions of binding sites are calculated with confidence levels from evolution and geometry information using machine learning algorithms.

Representation of protein 3D structure

Protein structures are represented in multi-scales. Basically, there are two levels of representation. At a higher resolution level, the protein structure is represented as a set of points with atom coordinates and van der Waals radius. The atoms are filtered out from the point set if they are (1) H atoms or backbone N/O atoms, (2) relative solvent exposed ratio is above 0.95, or (3) adjacent with atoms that are all buried. This method is referred as the "double smooth shell" representation of the protein structure. At a lower resolution level, each of residues in a protein structure is condensed as a point with 0 radius. The coordinates of the point are the mass center of Cα plus side chain atoms. These two representations will be applied to different stages in the algorithm. There are several advantages to apply this multi-scale representation. First, it will speed up the computation. Especially, time spent in regular triangulation step will be greatly reduced. It has the time complexity of $O(n^3)$, where n is the number of points. Second, it will improve the detection of the shallow pockets. Third, it will reduce the effect of side-chain placement on the prediction. Thus, the performance of the SiteSeeker will not be sensitive to low-resolution structure or homology models.

Determination of conservations of residues

The residue conservation is calculated using position specific amino-acid frequency (PSAF) based evolutionary trace method []. In brief, PSI-Blast [] is firstly used to search for homologous of the query sequence. The hits are filtered with the length criteria and HSSP sequence identity-sequence length curve []. The remaining homologous sequences are multiple aligned using ClustalW []. Position specific amino acid frequencies of the alignment are calculated as the following steps:

The sequences are weighted using Henikoff & Henikoff scheme []. Observed amino acid frequencies are calculated for all columns. For a gap with the weight w, a frequency of $w*a_i$ is added to amino acid i, where a_i is the observed background frequency of amino acid I from SWISSPROT database.

The pseudo count at each column is calculated based on observed frequencies. The higher conservation, the smaller pseudo count. Conservation is calculated using physical based amino acid exchange score matrix [] with the following formula:

$$c = \mathrm{SIGMA}(f_i*f_j*s_{ij})$$

where fi, fj is the observed frequencies of amino acid i and j. sij is the exchange score between i and j.

The background frequencies of 20 amino acids at each column are calculated from 9-component Dirichlet mixture[], given observed frequencies and pseudo counts.

PSAFs for each column are calculated given observed, background frequencies and pseudo count using the following formula:

$$p_{ij} = N_c*f_{ij}/(N_c+B_c) +B_c*b_{ij}/(N_c+B_c)$$

The conservation scores are calculated given PSSM using the following entropy formula:

csv = -SIGMA(p_i*log(p_i))

Finally, the conservation scores of all columns are normalized. The conservation score for a pocket is defined as the average of conservation scores of all residues in the pocket.

The local shapes of amino acids in the protein structure

Surface residues of a protein are defined as those residues with a relative solvent accessible surface area larger than 0.05. The convexity of the surface residue is computed with three steps. Firstly, the surface normal pointing to the solvent of each surface residue is calculated. In current implementation, the surface normal is represented with the solvent vector proposed by Jones & Thornton [Jones, S & Thornton, J. M. , JMB(1997)272, 121-132]. In brief, the residue in the protein is represented by its C_α atom. For a surface residue, the center of the mass of its 15 most nearest Ca atoms is calculated. The vector from this center of mass to the center of C_α atom of the surface residue is defined as the solvent vector of this residue. Compared with other method, solvent vector is a fast and robust way to define the local orientation of the protein surface. It also smoothes out the irregularity of the protein surface. Secondly, the surface neighbors of each surface residue are defined. Two surface residues are considered as the neighbor if their side chains contact each another. Finally, convexity of the surface residue is calculated by the ratio of number of residues up and below the plane defined by the center of Ca atom and the surface normal above it.

The rim and valley residues

The rim is defined as the residue that is most unlikely to be involved in the binding sites. On the contrast, the valley is the residue that may play a role in the binding. Two underlying assumptions are used here when determining rim or valley residues. First, evolutional conserved residues are more likely involved in the binding. Second, most of the binding sites are concaved. Thus, a residue is considered as a rim if it is less conserved or convex.

To decide whether a residue is a rim or a valley, it is critical to define both of the evolution and the geometry threshold. In the common practice, these kinds of the threshold can be refined from a set of training data. Unfortunately, due to the inherent diversity of protein surfaces and biased nature of sequence data, there is no such a set of universal parameters that fit to all cases. Therefore, we developed a self-adjusting parameter estimation algorithm based on information theory. The details of the algorithm will be presented in another report. In brief, the evolution threshold is estimated from the diversity of homologous sequences in the alignment. The more diverse are sequences included in the alignment, the more information content does it provide. The geometry threshold is estimated from the distribution of local shapes of surface residues. For example, an approximate uniform distribution implies that there are no obvious concave or convex regions on the surface. On the contrary, peaks in the distribution give clues on the size of concave regions. Thus, the evolution and geometry thresholds can be defined from the analysis of the distribution curve. The rim and valley residues are accordingly determined.

Delaunay Triangulation, alpha complex and surface mesh

Weighted Delaunay triangulation of the "double smooth shell" representation of the protein structure is calculated using the quick hull algorithm [] with the robust predicator []. Alpha complex are performed using Edelsbrunner's algorithm []. After the alpha complex is computed, it is easy to determine the surface meshes that are a set of triangles from the alpha complex to cover the protein surface. The “double smooth shell” representation significantly reduces the number of points required for the Delaunay triangulation and the surface mesh computation, but without lose of the surface information.

Construction of disjointed surface patches

A disjointed set of surface patches is constructed from the rim residues and the mesh triangles. Based on information from the rim residues, the mesh triangles can be classified as rim or valley triangles. A triangle is a rim triangle if it has more than one rim vertex. Otherwise, it is a valley triangle. A surface patch is constructed by join adjacent valley triangles. Two triangles are adjacent to each another if they share two vertices.

Construction of disjointed 3D pockets

After the surface patches have been determined, it is straightforward to union tetrahedrons of alpha complex into a disjointed set of pockets. Two tetrahedrons are union together if they share a common triangle with each other and in the same surface patch. For each pocket, its geometry properties such as volume, surface area, depth, and mouth area can be analytically calculated []. Furthermore, the pocket can be fitted into a generic surface type such as paraboloid, ellipsoid, and cylinder etc. and be further classified.

Estimation of Prediction confidence

The confidence of the site prediction is estimated using machine learning approaches. Here, support vector machine (SVM) is used to train and predict the confidence of the predicted binding site. The training data are 100 protein structures with known ligands bound, which are randomly chosen from the PDB []. The known ligand binding sites are defined as the true cases, and random generated pockets in these proteins are defined as the false cases. Three features are used in the SVM training and prediction: the conservation score, the volume and the depth of the pocket. A radial-based function is used in the training and the prediction.

2.2. Benchmarks

Two data sets, which represent respectively protein-ligand and protein-protein binding sites, have been compiled from public databases and literatures. The data set for protein-ligand binding sites is built from protein chains with known 3D structures of protein/ligand complex. Only small organic molecules and peptides are considered as the ligand. DNA/RNA molecules and metals are excluded. The PDB chains are selected from PDB SELCET with sequence identities less than 90%[]. Only x-ray structures are included in our test data set. There are total of () chains and () sites with the ligands in the final benchmark. The second one consists of protein chains with the protein-protein binding sites from 70 various types of proteins []. They include enzyme complex, G-protein, protease complex, protease-inhibitor, antibody-antigen and other miscellaneous proteins. After filtering redundant chains and convex binding sites such as antigens, there are 59 chains and 90 binding sites in the final data set.

Residues involved in the ligand or protein binding are derived from the change of exposed solvent accessible surface areas between complex and apo-state. Solvent accessible surface area is calculated using the algorithm proposed by [].

2.3. Performance Evaluations

The accuracy of the prediction is defined as the ratio of the number of the correct predicted residues and the total number of residues in the reference binding sites. However, the over-predicted binding site is less meaningful even if it has 100% accuracy.

Therefore, only the predicted site with the number of residues less than 120% of the referenced site is considered as the correct one. SiteSeeker was compared with PASS []. Performances are measured with two criteria: residue accuracy and rank accuracy. The residue accuracy is used to measure the goodness of correctly identified binding sites. The residue accuracy is measured with sensitivities to identify a binding site correctly. Here, the sensitivity is defined as a ratio of the number of correct predicted residues divided by the number of total residues in the true binding site. The rank accuracy is the rank of correctly predicated site in the all predicted sites of a protein.

3. Results and Discussions

3.1. Location Accuracy

SiteSeeker performs much better than PASS in identifying both protein-ligand and protein-protein interaction sites. As shown in Figure 2a, around 80% known protein-protein sites in the benchmark can be identified by SiteSeeker with sensitivity larger than 50%. At the same sensitivity level, PASS only identified around 20% of them. Moreover, SiteSeeker does not sacrifice its specificity to achive high sensitivity. The figure 2b shows that the specificity of SiteSeeker is slightly higher than that of PASS.

3.2. Rank Accuracy

There are more correctly predicated sites at top rank in SiteSeeker than in PASS, as shown in Figure 3.

3.3. Robustness of SiteSeeker

Another feature of SiteSeeker is that it is less sensitive to the size of the binding pocket. There is no big difference between large and small pocket in the overall prediction sensitivity and specificity by SiteSeeker . On the contrary, PASS performs much better in the large pocket than in the small pocket, considering both sensitivity and specificity.

3.4. Case studies

SiteSeeker has been successfully applied to real situations in the drug discovery processes.

3.5. Limitations of the algorithm

In the current implementation of SiteSeeker, only evolution and geometry properties are used in the prediction and evaluation. Another important surface property directly related to the ligand binding – surface physiochemical characterizations, such as electrostatic, hydrophobic fields, hydrogen-binding density etc. is not considered. Well-characterized physiochemical properties are easily incorporated into the current SiteSeeker framework, and will be expected to improve its performance significantly.