Analysis and Discrimination of Ligand Binding Membrane Proteins Using a Simple Statistical Approach

M. Xavier Suresh  M. Michael Gromiha  Makiko Suwa
xavier-suresh@aist.go.jp  michael-gromiha@aist.go.jp  m-suwa@aist.go.jp

Computational Biology Research Center (CBRC), AIST, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan

Keywords: membrane proteins, sequence analysis, prediction, ligand binding

1 Introduction

Many functionally important activities such as signal transduction, ion transport, and cell-cell communications, are mediated through an important class of proteins known as membrane proteins. Though studying them is difficult experimentally, in recent years the numbers of entries are rapidly increasing in the sequence databases whereas structural details are limited. In order to verify functional assignments, one must examine the common features inherent there with. In an attempt to identify the functionally important sites, we performed sequence and structural analysis on the interactions between the membrane proteins and ligands that would help to derive structural principles in protein-ligand complexes which provides deep insights in to the mechanism of protein-ligand interactions and their function. Therefore we have analyzed the membrane protein-ligand complexes through contacts between protein and ligand, accessible surface area, secondary structure, and location of the residues, etc. which in turn paves way to devise a simple statistical model to predict the ligand binding membrane protein sequences from the pool of membrane protein sequences with better accuracy.

2 Methods

2.1 Dataset

For the current study the protein structures were obtained from PDBTM [1], a database of transmembrane proteins derived from the protein data bank (PDB). Of the 661 transmembrane protein structures, excluding the NMR structures, fragments and repeated structures, 170 complexes are found to be non-redundant and complexed with various ligands. Based on ligand binding they were classified into binding and non-binding set. Similar sequences from the binding set which were present in non-binding set were removed. These two sets were once again checked for 90% identity to remove any redundant sequences.

The amino acid composition and the frequency of occurrence of the dipeptide and the motifs of various orders were computed and have been subjected to further analysis. The amino acid composition in a given sequence is calculated using the relation $\text{Comp}(i) = \frac{\sum n(i)}{N}$, where $n(i)$ is the number of amino acid residues of each type $i$ and $N$ is the total number of amino acid residues.

2.2 Prediction Method

We have calculated the amino acid composition for both binding set (CompBind) and non-binding sets (CompNon). For a new protein, X, firstly, we calculated the amino acid composition using the number of amino acids of each type and the total number of residues. Then we calculated the total absolute difference of amino acid composition between protein X and the amino acid composition of binding set, and that between protein X and non-binding sets. The protein X is predicted to be a ligand binding membrane protein if the deviation is lowest with CompBind, otherwise non-binding. Similar procedures with the dipeptide and motif compositions have been carried out. Similar approaches were successfully used earlier for discriminating beta barrel proteins [2] and DNA binding proteins [3].

Self consistency test and leave one out cross validation tests such as Jackknife test have been employed on the method developed. Overall prediction accuracy (PA) is estimated by using the following expression, $PA = \frac{(TP+TN)}{(TP+FP+TN+FN)}*100$, where TP, TN, FP, FN are true positive, true negative, false positive and false negatives respectively.

3 Results and Discussion

3.1 Sequence Analysis:

Earlier analysis of the atom contact on 170 membrane protein-ligand complexes shows that 7.23%, 27.91% and 64.86% of atom contacts are around the ligands at the vicinity of 0-3.0, 3.0-4.5 and 4.5-6.0 Å which reveals the fact that, in addition to the direct interactions, the influence of the neighboring atoms and residues...
may also play key roles in the binding of these ligands at the binding site [4]. The average deviations of amino acid composition of the binding and non-binding set of sequences are depicted in Fig. 1. The composition of charged residues (D, E, K, R) are higher in non-binding data set where as the composition of hydrophobic residues (A, G, L, M, V) are higher in ligand binding membrane protein sequences. Compositional preference of Isoleucine (I) is predominant in binding set whereas Serine (S) is predominant in non-binding dataset.

We have also analyzed the dipeptide and motif information. Analysis shows that 8.71% of dipeptides are in contact while it is 13.43% for single amino acid contact. Only 3.25% dipeptides occur more than 50%. Analysis of the neighboring residues indicates that the role of Cys is predominant. The occurrence of His and Met are more in highly preferred dipeptides, which dictates that these residues may have great affinity for ligand binding in membrane protein-ligand complexes. The knowledge thus gained is used further in predicting the ligand binding membrane sequences.

3.2 Prediction

Table 1 summarizes the prediction results using various parameters. The performance of the method over external dataset can be achieved by reduced features for the training. The prediction accuracy was fine tuned by the choice of parameters (for example 10 amino acid parameters give the accuracy 94%). The overall accuracy obtained was 81%. The choice of serine alone be able to predict with an accuracy of 73%, which shows that serine is an important amino acid that can mediate ligand binding in membrane proteins. Next to amino acid composition, predictions using the motif parameters perform with accuracy 83% (A****B, * being any amino acid), and was better than dipeptide composition. Validity of the method is examined by subjecting it to self consistency test and Jackknife test. During Jackknifing the overall prediction accuracy attained was 79%. The method performed well with out much deviation in the accuracy, there is room for further improvement of accuracy though.

4 Conclusions

Our method was able to discriminate the ligand binding membrane protein sequences with an overall accuracy of 83% from that of the non binding sequences. Various validity checks are also employed reveals that >79% prediction accuracy. The method with distinct parameters would have been implemented as an online tool and publicly available. Further this approach will be extended to identify the potential ligand binding sites which will effectively be used in function annotation and aid in drug design initiatives.

Acknowledgment

MXS is a Post Doctoral Fellow of JSPS, Japan.

References